QUALITATIVE DATA SHARING PRACTICES IN SOCIAL SCIENCES

by

Wei Jeng

B.A. in Library and Information Science, National Taiwan University, 2009

Mater in Library and Information Science, University of Pittsburgh, 2011

Submitted to the Graduate Faculty of

School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Wei Jeng

It was defended on

January 12, 2017

and approved by

Jian Qin, Professor, Syracuse University

Sheila Corrall, Professor, University of Pittsburgh

Liz Lyon, Interim Doreen E. Boyce Chair, University of Pittsburgh

Jung Sun Oh, Adjunct Assistant Professor, University of Pittsburgh

Dissertation Advisor:

Daqing He, Professor, University of Pittsburgh

QUALITATIVE DATA SHARING PRACTICES IN SOCIAL SCIENCES

Wei Jeng, PhD

University of Pittsburgh, 2017

Social scientists have been sharing data for a long time. Sharing qualitative data, however, has not become a common practice, despite the context of e-Research, information growth, and funding agencies' mandates on research data archiving and sharing. Since most systematic and comprehensive studies are based on quantitative data practices, little is known about how social scientists share their qualitative data. This dissertation study aims to fill this void.

By synergizing the theory of Knowledge Infrastructure (KI) and the Theory of Remote Scientific Collaboration (TORSC), this dissertation study develops a series of instruments to investigate data-sharing practices in social sciences. Five sub-studies (two preliminary studies and three case studies) are conducted to gather information from different stakeholder groups in social sciences, including early career social scientists, social scientists who have deposited qualitative data at research data repositories, and eight information professionals at the world's largest social science data repository, ICPSR. The sub-studies are triangulated using four dimensions: data characteristics, individual, technological, and organizational aspects.

The results confirm the inactive data sharing practices in social sciences: the majority of faculty and students do not share data or are unaware of data sharing. Additional findings regarding social scientists' qualitative data-sharing behaviors include: 1) those who have shared qualitative data in data repositories are more likely to share research tools than their raw data; and 2) the perceived technical supports and extrinsic motivations are both strong predictors for qualitative data sharing. These findings also confirm that preparing qualitative data sharing packages is time- and labor-

consuming, because both researchers and data repositories need to spend extra effort to prevent sensitive data from disclosure.

This dissertation makes contributions in three key aspects: 1) descriptive facts regarding current data-sharing practices in social sciences based on empirical data collection, 2) an in-depth analysis of determinants leading to qualitative data sharing, and 3) managerial recommendations for different stakeholders in developing a sustainable data-sharing environment in social sciences and beyond.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF BOXES

LIST OF DATA TABLES

ACKNOWLEDGEMENT

This dissertation would not have been possible without the support and assistance from many great people. Firstly, I wish to thank all the participants in this dissertation study. Their help is essential to the completion of this dissertation.

I would like to express my very great appreciation and gratitude to my advisor and dissertation chair, Dr. Daqing He, for his continuous guidance, inspiration, encouragement, and selfless support. His confidence in me motivated me to live up to my full potential, not only on my course toward the Ph.D. degree but also the journey afterward. My grateful thanks are also extended to faculty members who served as my committee, and to other professors who helped shape this dissertation throughout various stages:

- to Dr. Liz Lyon, for showing me the new world of data and being my muse;
- to Prof. Sheila Corrall and Dr. Jian Qin for providing insights which enriched this research; most importantly, for keeping me on the right track;
- to Dr. Jung Sun Oh and Dr. Brian Beaton, for the early supports and for being an inspiration on how to make work 'looks like work'. Without them, my academic interests are incomplete;
- to Dr. Stephen Griffin, for setting up a resourceful infrastructure and offering wise inspiration;
- to Dr. Eleanor Mattern, for being a great research partner and sweet friendship;
- to Dr. Steven Miller, for pointing the way out;
- to Dr. Chi-Shiou Lin and Dr. Hsu-Chun Hsiao, for pointing the way back.

In my earlier research projects, I have been opportunities to work with many brilliant faculty members at SIS. In particular, I would like to thank Dr. Leanne Bowler, Dr. Peter Brusilovsky, Dr. Kostas Pelechrinis, and Dr. Yuru Lin, for their guidance and creating an enjoyable collaborative environment.

I would also like to show my sincere thanks to the funding agencies that supported my PhD studies. These financial supports allowed me to follow my intellectual curiosity with few financial worries. I would like to thank the internal fellowships at SIS and the Government Scholarships for Study Abroad (GSSA) funded by the Taiwanese Government at the beginning of my doctoral program. I would also like to express my sincerest appreciation to the iFellowship, guided by Committee on Coherence at Scale (CoC) for Higher Education, sponsored by Andrew W. Mellon Foundations, which was not only a massive help adding to my doctoral studies, but also connects me to the LIS and iSchool communities. I also received funding assistance from the Eugene Garfield Doctoral Dissertation Fellowships by Beta-Phi-Mu Honor Society, which was incredibly beneficial in my last two semesters.

Many thanks to the excellent staff at SIS, especially Debbie Day, Wesley Lipschultz, Brandi Belleau, and Kelly Shaffer, for their availability and assistance. I was extremely fortunate to have many talented PhD colleagues and friends, to whom I am deeply indebted:

- to Sun-Ming Kim "Oppa," for being a supportive friend and reliable life mentor;
- to Jessica Benner, for shining my cloudy days in Pittsburgh and even in Seattle, and for the Bollywood dances in elevators of course;
- to Yu Chi, for 😂 (a.k.a. the face with tears of joy) and every little joy;
- to Shih-Yi Chien "James," for sharing time and everything.

I would also like to extend my thanks to my great PhD colleagues as research collaborators, Shuguang Han, Jiepu Jiang, Xidao Wen, Di Lu, Lei Li, Spencer DesAutels, Danchen Zhang, Rui Meng, and other SIS PhD fellows and iRiS Group fellows, who enriched my research and life. My special thanks to Joelle DesAutels, an excellent editor and gatekeeper, for her extensive editorial contributions that miraculously revived my writing.

I am forever grateful to my parents, Yu-Tin and ST, who offer timeless wisdom and inspiration, paved the way for me, believe in me with patience; and to my younger brother Eric who gently offers unconditional love and support.

1.0 INTRODUCTION

This dissertation investigates social scientists' qualitative data sharing practices, which have been under-investigated by previous work. Guided by two pre-existing conceptual frameworks, Knowledge Infrastructure (KI) and the Theory of Remote Scientific Collaboration (TORSC), this dissertation comprises two preliminary studies and three case studies. While the two preliminary studies paved the way for the design of the main study, the main study comprises three case studies, each of them aiming to 1) investigate the landscape of data-sharing practices in social sciences via the data sharing profile approach; 2) study the determining factors of participants' qualitative data-sharing behaviors; and 3) examine the world's largest social science data infrastructure's practices when curating and processing social science data.

This chapter overviews the research background and raises the research challenges of this dissertation study. It further defines the scope of this dissertation and identifies the research questions.

1.1 OVERVIEW

Sharing information, ideas and resources has always been recognized as a fundamental feature of scholarly collaboration and scientific discovery (Franceschet & Costantini, 2010). Among these sharable resources, research data has become a valuable cornerstone that allows scholars to make sense of inquiries, gain insights from evidence, develop humanity, and explain the world (Corti, Van den Eynden, Bishop, & Woollard, 2014).

Sharing research data has several immediate and long-term benefits. At an individual study level, research data sharing not only assists collective efforts to resolve complex research problems, but also facilitates the reexamination and enhancement of existing scientific theories and models. For researchers and their institutions, data sharing may increase visibility, opportunities, and scholarly impacts. Shared research data can also be utilized as teaching and learning resources that help train and educate the next generation of researchers, refine research methods, and advance science (Corti et al., 2014).

The recently-released mandate from the National Science Foundation (NSF) illustrates this data-sharing need; the mandate requires that all grant submissions, after January 18, 2011, include a supplemental "Data Management Plan" (hereafter: DMP). Entities affected by this policy shift include social-science-related directorates and allied units: the NSF Directorate for Social, Behavioral & Economic Sciences (SBE), Education & Human Resources (EHR), and the Institute of Education Sciences (IES).

Besides funding agencies, academic organizations in social science domains increasingly demand that scholars present their research evidence and ensure the openness of their data (Elman & Kapiszewski, 2013; 2014; APSA2012), demonstrating the acceptance of a common position on data sharing. For example, in October of 2012, the American Political Science Association (APSA)

revised *A Guide to Professional Ethics in Political Science* in order to reflect new requirements that encourage scholars to do their "best to ensure that no restrictions are placed on the availability of evidence to scholars or on their freedom to draw their own conclusions from the evidence and to share their findings with others" (APSA, 2012; Lupia & Elman, 2014). Another example can be seen in the American Anthropological Association (AAA)'s "Code of Ethics," which suggests that "[r]esults of anthropological research should be disseminated in a timely fashion" (AAA, 2012). Given the recent mandates from institutions, publishers, and funding agencies, as well as the encouragement from professional associations for data management and sharing plans (ROARMAP, 2016), sharing data has become a movement, an expectation, and also common sense.

However, previous studies have revealed that researchers are often reluctant to make their data available to others. Reasons for this reluctance include: insufficient time, too much effort, perceived risks such as fear of data misinterpretation and misuse, few perceived returns, and lack of incentives (Tenopir et al., 2011; Kim, 2012). The same barriers also plague social scientists. Worse yet, those who conduct qualitative studies can face additional challenges due to the different nature of qualitative data, the unique norms of social science, and lack of supports.

Different nature of data. First, sharing qualitative data is fundamentally different from sharing quantitative data due to the complexity and context-dependent nature of the former (Tsai et al., 2016). Qualitative data is complex because it has diverse data types and most are loose-structured (e.g., text-heavy). It is difficult to organize the data in a pre-defined table or database. Qualitative data is context-dependent because qualitative research usually involves individuals within a system or a society. Therefore, sharing and reusing qualitative data relies upon thorough context documentation, which requires much more effort.

Unique research norms. Social science research often deals with human society and relationships between individuals. This requires that social scientists take extra ethical considerations regarding

their studies. These ethical considerations include the protection of study participants and the clarification of the proprietary rights over data (Cliggett, 2013). These extra considerations can sometimes hinder researchers from sharing their qualitative data and results in the lack of strategic planning for long-term preservation.

Limited supports. Finally, social scientists who deal with qualitative research data face critical infrastructural issues, such as the lack of equipment, access, funding, and investment in infrastructure (Corti, et al., 2014; Prescott, 2013; Elman & Kapiszewski, 2013). These infrastructural and financial barriers impede qualitative researchers in social sciences from embracing more robust modes of data sharing.

Qualitative approaches have been widely adopted in many social science areas. Recent studies examine the presence of qualitative studies in core journals and conferences in linguistics and educational and information behavior, revealing that approximately 40% to 70% of articles are based on qualitative approaches (Benson et al., 2009; da Costa, 2016; McKenzie, 2008). Despite the presence of qualitative studies in social sciences, there is no systematic study to comprehensively identify the factors that influence such studies and their relationships between each other. This dissertation study aims to fill this void.

1.2 RESEARCH BACKGROUND

1.2.1 The era of e-Research

The origin of the research data management issue can date to the e-Research movement in the 2000s. The predecessors of e-Research are cyber-infrastructure and e-Science, terms coined in the early 2000s to highlight the importance of information technology that supports scholarly activities. According to Borgman (2007), the United States uses the term "cyber-infrastructure," whereas Asia, Europe, Australia, and other areas favor the term "e-Science." The prefix "e" in e-Science is usually taken to stand for "electronic," but can also be understood as "enable" or a concept of "enhancement" (p.20).

E-Research is often viewed as an extension of e-Science and cyber-infrastructure, incorporating e-Humanities and e-Social Sciences (Borgman, 2015). The Association of Research Libraries (hereafter: ARL) describes e-Research as a concept that encompasses "computationally intensive, large-scale, networked and collaborative forms of research and scholarship across *all* disciplines" (ARL, n.d., para 1). The scope of all disciplines, as ARL suggests (n.d.), includes "all of the natural and physical sciences, related applied and technological disciplines, biomedicine, social science, and the digital humanities" (para 1).

Consequently, e-Research describes research activities or its development programs as taking place in a Web-based environment, which usually generates a large amount of data and requires better research data management. Given that Hey and Trefethen (2003) foresaw the "Data Deluge" having "profound effects" on current scientific infrastructure, research data management and its related topics have emerged in e-Research's agenda (Hey, Tansley, & Tolle, 2009).

1.2.2 Digital scholarship and data scholarship

According to Unsworth (2006), *digital scholarship* is a set of scholarly practices geared toward 1) building a digital collection of information; 2) studying digital information, objects, and cultures; 3) conducting studies throughout the research lifecycle in a digital medium; and 4) creating tools, services, and resources for supporting research in the digital environment. The relationship between e-Research and digital scholarship is that e-Research (or cyber-infrastructure) describes a research environment built with digital structures and facilities, whereas digital scholarship emphasizes incorporating emerging digital supports to ensure that intellectual products can be accessible, disseminated and co-produced.

Griffin (2015) comments on contemporary digital scholarship by mentioning its basic characteristics: "rich dialog, shared and open access to resources and an emphasis on transparency." Here, one can see that data plays a very special role in the support of digital scholarship, because it provides the base resource of research, and enables research transparency, and communication between scholars.

While digital scholarship emphasizes supporting technology for research, *data scholarship* was referred to as "data-intensive research" in the 2000s (Borgman, 2015). Data-intensive research involves a broad range of scholarly activities, including computational analysis and a combination of many sources across multiple disciplines. Broadly speaking, data scholarship can also describe the complex relationship between scholarship and data. Boyer (1990) has indicated the general view of scholarship: discovery, integration, application and teaching. Borgman (2007) added *data* as another aspect to the concept of scholarship.

One shared concept of digital scholarship and data scholarship is that they both acknowledge the importance of data in research in the digital environment, which forms the background of this dissertation study.

1.2.3 Demands for research data management

In the discussions of e-Research movements in the 2000s, many scholars conclude that the explosion of scientific data has led to increasing computation requirements. More plans, controls and management are needed to face the "Data Deluge" and advance data scholarship. In response to the popularity of e-Research and data scholarship, the NSF has held a series of professional e-Research workshops and conferences (Friedlander, 2009). Since 2002, the NSF has engaged in organizing councils and digital scholarship workshops, producing several high-impact reports, including *Cyberinfrastructure Vision for 21st Century Discovery and Understanding Infrastructure: Dynamics, Tensions, and Design* in 2007. This series of movements and endeavors reflects the government's view on research data management: the data deluge requires more control over data management. Later, the U.S. government announced a manifesto of digital stewardship in 2009 and preannounced a mandate in 2010 that all NSF applications should include a research data management plan.

Based on this preannouncement, all NSF grant applicants, on or after January 18, 2011, are required to submit a two-page research data management plan describing how to share and manage their data. U.S. federal funding agencies further expanded this mandate in 2013 by adding new data management and data-sharing requirements to grant applications.

Besides the NSF, other major funding agencies such as the NIH (National Institutes of Health) and NEH (National Endowment for the Humanities) also published research data management mandates (Halbert, 2013). The NIH has long required sharing research data; as early as

2003, the NIH has promoted a data-sharing mandate (Goben & Salo, 2013). The NEH also has a statement that their policy is aligned with the NSF (NEH, n.d.).

Mandates from funding agencies change scientists' behaviors. Diekema, Wesolek, and Walters (2014) administered online surveys to STEM faculty members and discovered that the majority of faculty (56.8%) already stored or shared their data even before the NSF/NIH mandates; 25.52% of participants in the survey stated that they have changed their behaviors due to the mandates (Diekema et al., 2014). However, despite the popularity of e-Research and the NSF/NEH mandates, there is a particular absence of studies that focus on qualitative data sharing in social science disciplines.

The NSF's mandate on data management has also become a source to explore how PIs share and reuse their data. Mischo, Schlembach, and O'Donnell (2014) analyzed 1,260 DMPs from July 2011 to November 2013 at the University of Illinois. They found that the most common venues used by PIs to preserve their datasets were personal websites (40%), personal servers (42%), local institutional repositories (e.g., IDEALS at UIUC, 53%), and repositories that are not located on campus, including disciplinary repositories (22%) and other non-UIUC organization (28%). Among all 1,260 DMPs, the authors calculated the occurrence of named repositories mentioned by PIs. The arXiv, GenBank, and NanoHub are among the most frequently mentioned. However, the project of Mischo et al. did not find significant differences in storage venues when comparing funded grants to unfunded proposals. Additionally, they found that NSF grant applicants underutilized disciplinary repositories.

1.3 RESEARCH MOTIVATIONS AND QUESTIONS

This dissertation is motivated by the desire to understand social scientists' qualitative data-sharing practices, and such research inquiries include: 1) the current landscape of qualitative data sharing practices; 2) whether funder data-related mandates can fit both qualitative and quantitative data, and 3) whether the best practices or sharing strategies for qualitative data exist.

However, few empirical studies have been conducted to probe into the above research inquiries in the context of qualitative data sharing (Karcher, Kirilova, & Weber, 2016). Are qualitative data shareable? How do social scientists share qualitative data? What kind of reasons do researchers have for sharing or not sharing data?

To address these questions, this dissertation study formulates and answer two central research questions in order to unveil research data-sharing practice from both generic and focused perspectives.

Research Question 1 (RQ1): What are social scientists' general data-sharing practices?

Given the broad scope of RQ1, the following four sub-questions are raised, the outcomes of which, in combination, help to answer RQ1:

- **RQ1A:** What data (e.g., types of sources, format, and size) do social scientists interact with through different stages of their research processes?
- **RQ1B:** What are social scientists' current data-sharing practices (e.g., frequency and sharing channels)?
- **RQ1C:** What are the perceived community practices regarding data sharing in social sciences?
- **RQ1D:** What are the underlying technologies or other resources supporting data sharing in social sciences in the social scientists' work environment?

The first research question (RQ1) considers research data-sharing practices in a general context; that is, the scope is not limited to the scope in qualitative studies. The outcome of RQ1 can help identify whether barriers or incentives exist in qualitative or quantitative studies or both. This funnel approach (i.e., from a broad research question to a narrow interest) allows for the research environment to be scanned first to establish a bond of common practice and knowledge within the research topic.

The second question focuses on the determining factors of qualitative data sharing:

Research Question 2 (RQ2): What are the factors influencing qualitative data sharing?

Similarly, the RQ2 can be achieved by answering the following inquiries:

- **RQ2A:** What data do social scientists consider "shareable"?
- **RQ2B:** What are the factors positively influencing qualitative data sharing?
- **RQ2C:** What are the challenges of qualitative data sharing in social sciences in terms of community norms and underlying technological infrastructure?

Unlike RQ1, which captures generic data-sharing activities and practices from social scientists, RQ2 has a specific viewpoint which focuses on researchers with qualitative data-sharing experience, as well as data curation professionals who handle research data sharing and curation processes in a research data infrastructure.

While those basic empirical findings in RQ2 have been identified and carried out, this dissertation study also develop a coherent theoretical framework. The theoretical framework is developed to build greater understanding of the relationships among researchers' individual concerns, motivations, data characteristics, technological infrastructure, and research context in data sharing.

To answer these research questions, this dissertation study comprises three case studies. First, a preliminary instrument as a profile tool is used in Case Study 1 (hereafter: CS1) to collect

social scientists' data practices in order address RQ1. Data in CS1 were collected from 66 early-careered, currently-enrolled PhD students and post-doctoral students from the University of Pittsburgh and Carnegie Mellon University in the U.S. Based on CS1, a refined instrument is used in Case Study 2 (CS2) as a questionnaire, and sent to PIs who have shared qualitative data at the following research data repositories:

- Interuniversity Consortium for Political and Social Research (ICPSR), the world's largest primary data archive of social science research, and
- Qualitative Data Repository (QDR), the pioneer qualitative data repository in the U.S., hosted by Syracuse University

Case Study 3 (CS3) reports a study that comprises two focus group sessions and one individual interview with eight total employees at ICPSR.

The outcomes of this dissertation study include three parts: 1) descriptive facts regarding current data-sharing practices in social sciences, 2) an in-depth analysis of determinants leading to qualitative data sharing, and 3) managerial recommendations for different stakeholders in developing best practices for sharing qualitative data.

These outcomes are expected to advance the understanding of data-sharing practices in the social sciences, such that constructive suggestions can be provided to all parties, including researchers, academic libraries, and data repositories. The methodology design and theoretical framework, though developed for social sciences, can be also a starting point to assess the motivations and barriers regarding researchers' data-sharing practices.

1.4 SIGNIFICANCE

This dissertation examines data-sharing practices in fields outside of STEM, which have been thus far under-investigated. Given that data management and curation issues have recently received more attention in the library and information science and information science (hereafter: LIS/IS) community, the findings of this dissertation study can help information professionals become better designers, supporters, and consultants for social science data infrastructures. The findings also encourage outside agencies and organizations to focus more attention on the unique nature of qualitative data in social sciences.

On a continuum of data sources, social science disciplines exist in the middle ground between the STEM sciences and humanities (Borgman, 2009). An improved understanding of the sharing practices and needs of social science scholars will not only serve as a foundation to build more sustainable social science data infrastructures, but can also, more broadly, further data openness and collaboration.

Besides the contribution to the LIS/IS community and social science fields, the research findings and methods in this dissertation study could potentially be generalized and applied to other domains that produce qualitative data. These fields include, but are not limited to, arts, humanities, and behavioral sciences. More and more researchers have recognized the importance and effectiveness of using qualitative research methods in medical research (Borreani, Miccinesi, Brunelli, & Lina, 2004; Tong, Winkelmayer, & Craig, 2014) and other health sciences (Mori & Nakayama, 2013).

2.0 LITERATURE REVIEW I:

DATA-SHARING PRACTICES IN SOCIAL SCIENCES

This chapter serves two main purposes. First, it examines the definitions of several concepts in this dissertation study, such as research processes, research data, the realm of social science, and the definitions of qualitative studies and data. Second, it determines what has already been explored and established in the empirical literature about the nature of social-science research and data, and the challenges of qualitative data sharing.

2.1 RESEARCH & DATA IN SOCIAL SCIENCE

Research in the humanities and social sciences has a unique nature, centering on the protection of individuals and its methodological characteristics. For social science studies involving human participants, ethical behaviors guide the protection of individuals, communities and the environment (Israel, 2015). Researchers in the realm of *sociology of social scientific knowledge* have discussed how social scientists embody values and use their tacit knowledge when conducting survey research (Maynard &Schaeffer, 2000).

According to the Oxford Dictionary (n.d.), social science is defined as "the scientific study of human society and social relationships," and by Merriam-Webster Online (n.d.) as "a branch of

science that deals with the institutions and functioning of human society and with the interpersonal relationships of individuals as members of society."

In this dissertation study, social science is an umbrella term that encompasses these definitions and scopes: a set of academic disciplines concerned with human activities, social phenomena, and the relationships among individuals within a society. Possible social-science subjects, as the NSF Survey of Earned Doctorates (2014) suggests, include but are not limited to: anthropology, gender studies, political science & government, sociology, cultural studies, international relations, linguistics, urban studies, and economics. Disciplines listed as "NEC (not elsewhere classified)" but that fit in the definition are also considered social sciences, such as education, law, library science, social work, and public administration.

2.1.1 Research process

To better understand the role of research data sharing in social sciences, this section discusses where research data sharing occurs in the academic research process.

As shown in Table 2-1, even though the academic research process is often simplified as a linear model, most social science research involves a continual process composed of several activities such as designing, planning, and execution. Researchers also note that "[r]esearch is an iterative process of observation, rationalization, and validation" (Bhattacherjee, 2012, p. 20). This process guides a researcher to an outcome of their inquiries.

In general, social sciences can be divided into two methodological strands:

- Quantitative methods (post-positivism), wherein the researcher is motivated to validate a theory (i.e., deductive research); and

- Qualitative methods (constructivism), wherein the researcher starts at a phenomenon and attempts to rationalize observations (inductive research) (Abbott, 2001; Bhattacherjee, 2012).

Another strand, mixed methods (i.e., incorporating elements and characteristics of both quantitative and qualitative methods), is recognized in the field and represents the worldview of pragmatism (Creswell, 2009). The preference of qualitative and mixed methods reflects the worldview of many social science researchers: human behavior within a society is not an objective matter.

Most disciplines depict the general research process in a sequential order that reflect the "journey" of the research (Malins & Gray, 2013). On the one hand, the research process can be visualized as a graph whose nodes represent components and whose links indicate the order of occurrence. Depending on the graph's structure, research processes in social science research can also be visualized as a lifecycle or even a complex structure.

Table 2-1. Common research process patterns in humanities and social sciences

Category	Sub-category	Exemplar disciplines and studies	Summary of Characteristics
Linear	Linear	Qualitative studies in health science (Gómez, 2009); Social research in general (Bhattacherjee, 2012); Education (Fraenkel & Wallen, 2003)	In a linear process, every step depends on a sequential development. The endpoint has no arrow pointing back to the startpoint.
	-with subprocess	Business (Faisal, 2011)	A variance of the linear process: it may contain a subprocess that forms a cycle in one or more phases.
	Flowchart	Business (Sekaran, 2006); Business (Faisal, 2011)	A variance of the linear process: it contains flowchart elements such as decision (usually with a Y/N decision question)
Cycle	Cycle	Education (Johnson & Christensen, 2008); General (Leland Speed Library at Mississippi College, n.d.)	A cycle process might have a startpoint and an endpoint. However, some have no explicit startpoint and endpoint. For any node, one can go back to the same node by moving along the directed links.
	-with sub-cycle	Management (Viktor, 2008)	A variance of the cycle process, as a cycle process containing one or more smaller sub-cycles.
Hybrid	Daisy (or Star)	General scientific domains (Mackey, 2009; Mark & Helen Osterlin Library, n.d.) ; General (University of California Museum of Paleontology, 2008)	The research process can also form a "daisy" or a "star" shaped graph with the central idea placed in the center, connecting to neighboring components via links (often bidirectional). These neighboring components form a cycle among themselves, too. This structure allows high flexibility at visualizing the course of research or only one stage of research.
	Network	Behavioral science in general (Hayes, 1997). Art and Design (Malins & Gray, 2013)	The research process forms a complex network with one-directional or bidirectional links to any component on the graph. The components might have a sequencial order but they may also be interconnected.

While qualitative research processes might vary, common components can be condensed into four main areas: conceptualization, design, execution, and reporting (see Table 2-2). Note these four components are typical but not required, and there is no specific chronological order among them.

Table 2-2. Common research components in social science research

Themes	Example research activities	Studies
Conceptualization	Developing research questions	Bhattacherjee, 2012; Otago Polytechnic, 2006; Fraenkel & Wallen, 2003…
	Literature review	Bhattacherjee, 2012; Fraenkel & Wallen, 2003
Design	Selecting research method	Bhattacherjee, 2012; Johnson & Christensen, 2008
	Defining variables, samples	Fraenkel & Wallen, 2003
Execution	Measurement	Viktor, 2008
	Data gathering/collecting	Bhattacherjee, 2012; Otago Polytechnic, 2006; Fraenkel & Wallen, 2003…
	Data analysis	Bhattacherjee, 2012; Otago Polytechnic, 2006; Fraenkel & Wallen, 2003…
Reporting	Authoring	Bhattacherjee, 2012; Johnson & Christensen, 2008; Hayes, 1997; Gomez, 2009
	Presentation	Otago Polytechnic, 2006
	Publishing	Hayes, 1997

The research data lifecycle may serve as a sub-cycle, which often occurs during the execution and dissemination stages of a research process (University of Virginia, n.d.). The UK Data Service's research data lifecycle is adopted for this dissertation study, and assumes that data sharing occurs during the "giving access to data" stage (see Figure 2-1 below from the UK Data Service).

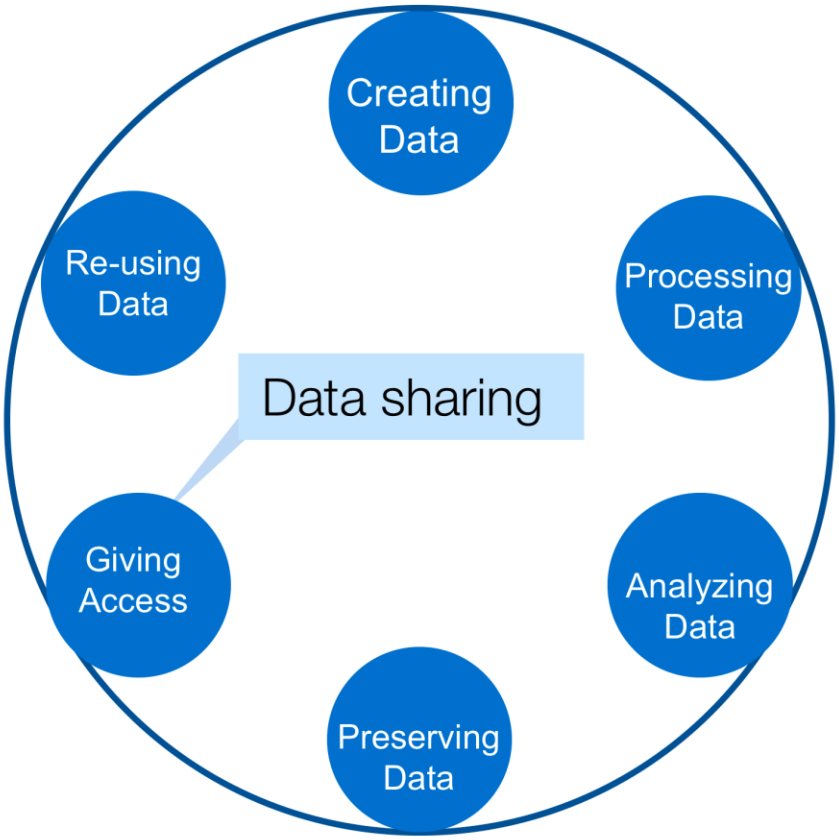


Figure 2-1. Data lifecycle

Source: Redrawn by this dissertation from UK Data Service (n.d.).

This setting entails the following two clarifications in the research scope.

First, "giving access to data" is within the research scope of this dissertation, while the two other stages, "data preservation" and "data reuse," become a supplemental background with less focus in this dissertation study. Data reuse and data preservation are very important because they are precisely close to the phase of data sharing. However, this dissertation study specifically focuses on data sharing and will discuss data reuse and presentation as needed.

Second, although this dissertation study defines the research scope of data sharing under one type of data process (i.e., lifecycle), one should note that there is much variation: data processes can be just as diverse as research processes.

2.1.2 Data in social sciences

The term "data" can be seen as early as the 18th and 19th centuries, making it dissimilar to the buzzwords "social media" and "cloud computing" (Borgman, 2015). Despite its long history, only recently has "data" become a popular research topic, for the reasons mentioned in Section 1.2.

The federal government defines research data (i.e., OMB Circular A-110) as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." "Research data" commonly refers to the raw material obtained or generated during the course of research work. There are three critical attributes of research data: when and where the data are obtained (situation), what they comprise (content), and why they are used (context) (Martin, 2014).

Concerning the term *context*, how do social science-related funding agencies define the meaning of data? The NSF Directorate for Social, Behavioral & Economic Sciences does not have a customized definition of data, instead following the federal government's definition. The Institute of Education Sciences (IES) clearly emphasizes the importance of raw data. According to their data-sharing policy, final analytic results (such as summary statistics or tables) are not data that researchers should share; instead, researchers should focus on "the factual information on which summary statistics and tables are based" (IES, n.d.). Therefore, according to IES (n.d.), "laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research,

peer review reports, or communications with colleagues" do not count as final analytic results and are not expected to be shared.

Another social- and humanities-related funding agency, the National Endowment for the Humanities (NEH) Office of Digital Humanities (2014), defines the term "data" as

> *"materials generated or collected during the course of conducting research. Examples of humanities data could include citations, software code, algorithms, digital tools, documentation, databases, geospatial coordinates (for example, from archaeological digs), reports, and articles."*

The NEH's guideline, compared with the NSF SBE's, is more tailored for the related disciplines. However, NEH also identifies data that do not count:

> *"things such as preliminary analyses, drafts of papers, plans for future research, peer-review assessments, communications with colleagues, materials that must remain confidential until they are published, and information whose release would result in an invasion of personal privacy (for example, information that could be used to identify a particular person who was one of the subjects of a research study)" (para 3).*

2.1.3 Norms in social sciences

Norms, ethics, and/or community-best practices influence the behavior and decision-making of social scientists. Historically, different disciplines, institutions, or professional communities guided and governed ethical statuses and practices of research (Haggerty, 2004). These norms suit each community's particular behavior and goals, which helps its members to "coordinate their actions or activities and to establish the public's trust of the discipline" (Resnik, 2010, para 6).

Israel and Hay (2006) conclude that for social scientists, ethical behaviors and research norms are put in place "[to] avoid doing long-term, systematic harms to individuals, communities, and environments, and offers the potential to increase the sum of good in the world" (p.2). Besides these examples, other fundamental research norms and ethics in social sciences include ensuring research integrity and satisfying peers and/or community demands (Israel, 2015). When applied to the context of research data sharing, the above norms and ethics are consolidated into two objectives. One is protecting others and minimizing harm, and the other is research integrity (Borgman, 2015).

Protect Others and Minimize Harm. For social science studies involving human participants, ethical behaviors oversee the protection of individuals, communities and the environment (Israel, 2015). While dealing with data, social scientists should guarantee the confidentiality of participants by using anonymization techniques. Researchers should also minimize harm to participants by avoiding "psychological, social, economical, legal and environmental damage" to the participants during studies (Israel, 2015, p.124).

Research Integrity for Social Scientists. To maximize the quality of disseminated information, NEH (n.d.) requires that grant awardees be "wholly responsible for conducting their project activities and preparing the results for public distribution unless specifically authorized to represent information on behalf of the agency" (para. 4). Moreover, the federal government also highlights the importance of honesty and accuracy in research (Office of Science and Technology Policy, 2000). These research norms, applied in the context of data scholarship, can be adhered to by sharing and submitting data for peer review. Since data sharing helps ensure research integrity, the academic community should act as a gatekeeper that carefully examines the accuracy of research information and datasets.

In summary, social science's research norms define how social science scholars view their data and participants in response to data-sharing requests. On the one hand, social scientists respect their participants and are concerned about potential privacy leakage due to data sharing. On the other hand, they may consider data sharing an enticing means to ensure research integrity. Thus, the tension between these two forces is the key factor affecting the decision to share data. More discussions on ethical considerations and challenges can be found in Section 2.4.

2.2 SOCIAL SCIENCE DATA-SHARING PRACTICES

2.2.1 Data-sharing practices before the Internet was commonly used

Social science researchers have been sharing data for many years. As early as the 1970s and 1980s, literature has documented that social scientists used others' data to verify original studies or to reanalyze and produce new research (Fienberg, Martin, & Straf, 1985). Gerry King, the director of the Institute for Quantitative Social Science at Harvard University, once described the political science community as needing existing data to verify and replicate others' study outcomes: "the community of empirical political scientists needs access to the body of data necessary to replicate existing studies to understand, evaluate, and especially build on this work" (King, 1995, p.444).

Not long after data-sharing practices began, advocacy for and concern about data sharing emerged. According to King (1995), data sharing among political scientists is always troublesome and nearly impossible because of the lack of solid documentation on studies.

Similar practices can be found in other social science domains. Investigators in comparative sociology need to compare their analyses on different data sets "in order to generalize findings about

social phenomena" (Fienberg et al., 1985, p.10). Fienberg et al. (1985) provided a list of data-sharing benefits, specifically in social sciences (p.124-p.130, extracted sub-headings):

- "*Reinforcement of open scientific inquiry*
- *Verification, refutation, or refinement of original results*
- *Promotion of new research through existing data*
- *Encouragement of new appropriate use of empirical data in policy and evaluation*
- *Improvements of measurement and data collection methods*
- *Development of theoretical knowledge and knowledge of analytic techniques*
- *Encouragement of multiple perspectives*
- *Provision of resources for training in research*
- *Protection against faulty data*"

In hindsight, social scientists in the 1970s and early 1980s did not distinguish between data digests from publications and raw data, as they believed that publishing academic papers was a type of data sharing. This definition, however, differs from what the federal government mandates today (see Section 2.1.2).

In the 1970s and 1980s, "data collecting facilities" usually referred to personal computers (called micro-computers at the time) and databases; data had to be transmitted using portable storage devices such as magnetic tapes, floppy and hard disks, cassettes, and so on (Clubb, Austin, Geda, & Traugott, 1985; Sieber, 1991). As for data storage sites, Sieber (1991) had already mentioned ICPSR, which was the top choice for data storage for many social, behavioral, and political scientists. Other than ICPSR, General Social Survey (GSS), now affiliated with the University of Chicago, was also mentioned in the 1980s.

2.2.2 Data-sharing practice in the digital age

The advancement of communication technologies has drastically changed how researchers share data. Cragin, Palmer, Carlson, and Witt (2010) studied scholars' data-sharing practices and willingness from small sciences in the 2000s. The scope of their study included humanities (e.g., history) and social science disciplines (e.g., linguistics, etymology, and sociology) in the Data Curation Profiles project (datacurationprofiles.org). This project reveals common standards for social science disciplines, such as preferred file exchange formats, file size, and preferred embargo time (i.e., data not published or shared until a set date or certain conditions have been met) for researchers. Based on the curation profile published on the project website (Zilinski & Lorenz, 2011; Tancheva, 2012), social science subjects such as linguistics and etymology handle very large files in a single project: each file can range from 150 MB to 3 GB. However, embargo time varies. One research team may set a 5.5-year embargo time, whereas another team may not specify their restriction. Cragin et al. found that very few scholars routinely deposit their research data into data repositories. The study also shows that although establishing resources and services for shared data is considered important, no field-wide norms have been established.

Compared to STEM disciplines, social science scholars are more likely to be concerned about data misuse by others. Tenopir et al. (2011) conducted a national survey that recruited 1,329 scientists, including 204 social science researchers. The survey found that social science researchers are less likely to make their data electronically available to others when compared with STEM scholars: only 47 out of 204 (23%) agreed or somewhat agreed that their data could be easily accessed by others. The percentage agreement from scholars in atmospheric science and biology were nearly two times higher or more (39% and 49%, respectively). Overwhelmingly, 162 out of the

204 social scientist participants (79%) in the survey agreed or somewhat agreed that they had concerns about data being used in unintended ways.

To sum up, the Internet and cloud storage technologies do help disseminate data. However, reviewing social science researchers' data-sharing practices in the 1980s and 1990s reveal that their data-sharing concerns and challenges do not significantly differ from ones today.

2.2.3 Social science data sharing in interdisciplinary domains

Researchers in different disciplines interact with distinct kinds of data they create and gather. During this process, there might be significant variations in their data-sharing needs, attitudes, and practices (Cragin et al., 2010). Since many research questions require interdisciplinary problem-solving in the social sciences, it is common for social science scholars to participate in cross-discipline collaboration and use data from other domains. For example, anthropologists integrate legal documents or medical records, and political scientists need data from ecological surveys.

Several research studies mention data sharing in interdisciplinary or cross-disciplinary scenarios, and their results suggest that social scientists have positive attitudes about interdisciplinary sharing. White (1991) describes anthropology's data sharing in a cross-disciplinary setting, with the example of anthropology combining the earth and environment using time-series remote sensing data.

Due to the variation in disciplinary (or sub-disciplinary) data practices, interdisciplinary data sharing can be difficult and thus requires additional management strategies. Parsons et al. (2011) provide four guidelines for interdisciplinary data sharing: data should be discoverable, open, always linked, and useful. These indicate that cross-disciplinary data sharing or research collaborations require better infrastructure (Lim, Iqbal, Yao, & Wang, 2010), including improved standards and

services, and more research about cross-disciplinary and interdisciplinary data-sharing challenges is needed.

2.3 DATA SHARING STANDARDS IN SOCIAL SCIENCES

This section reviews technical standards for data sharing in social sciences. This review uses a funnel approach to discuss two levels of technical standards that data sharing needs most: the *infrastructure level standard* (which frames a research data curation or archiving service), and the *metadata level standard* (which is applied to the data package).

2.3.1 Technical framework for the service level: the OAIS

One well-known infrastructure-level or organization-level standard is the Open Archival Information System (OAIS), which is an ISO standard for creating and maintaining a digital data repository. The OAIS can be understood as a framework that helps an "organization or system charged with the task of preserving information over the long term and making it accessible to a specified class of users (i.e., designated community)" (OCLC, n.d.). The OAIS was proposed two decades ago and has become a widely-adopted conceptual model for "maintaining digital information over the long-term" (Lavoie, 2004, p.2).

The OAIS model can be viewed at three different levels of granularity. The *macro level* describes the external world with which an OAIS interacts. According to Lavoie (2004), the external world of OAIS comprises three entities (Figure 1 in Lavoie, 2004): producer (the party who submits the information package for the OAIS to preserve), management (responsible for high-level policy framing work), and consumer (a.k.a., designated community, the party that interacts with or uses the

final outcome of the preserved archives). It is worth noting that the management entity is "not responsible for overseeing the day-to-day operations of the OAIS" (Lavoie, 2004, p.5). Such a responsibility is handled within the OAIS itself.

The *meso level* defines the internal workflow of OAIS, including six functional entities (Figure 2-2). After an information package is submitted by a producer as a submission information package (SIP), it continues to interact with each functional entity. Such an information package, also considered the *micro level* of OAIS, is converted from an initial SIP to an archival information package (AIP) and finally a dissemination information package. The micro level defines the format of possible inputs to the OAIS services.

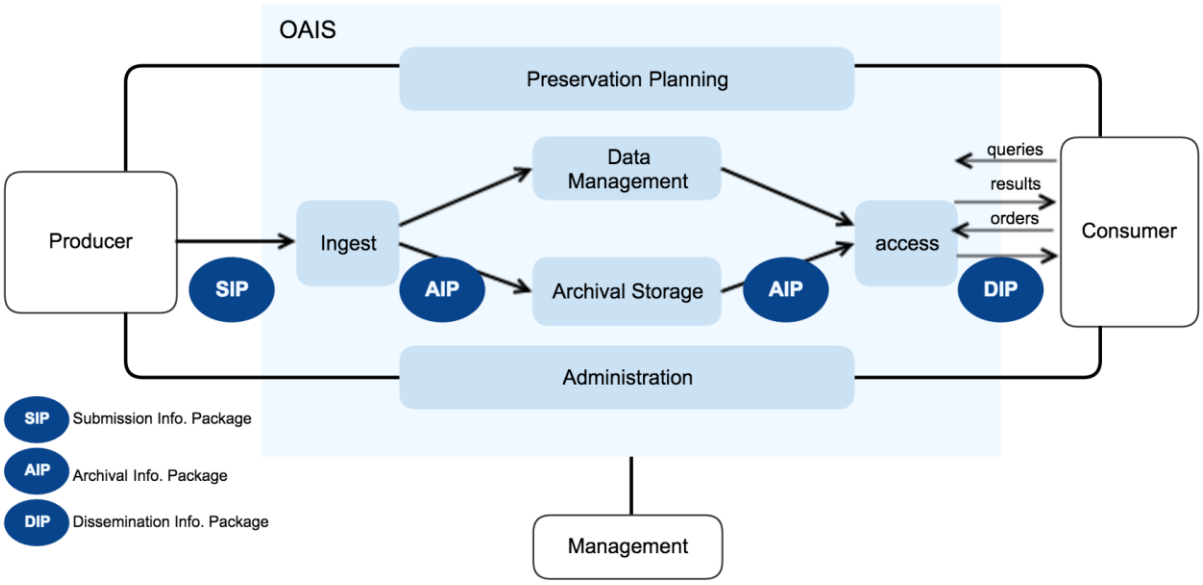


Figure 2-2. Meso level: Functional model of OAIS

Source: Redrawn by this dissertation based on Wikimedia Commons

Table 2-3 is used to summarize the interactions between the information packages (micro level), OAIS functional entities (meso), and external entities (macro). Starting with the potential data depositors ("producer" in Figure 2-2) submitting their data package (SIP) to the repository, the information package is processed via the *ingest* functional entity, producing an archival information package. Next, several functions are applied to this AIP, including *archival storage* (the functional entity that ensures that "archived content resides in appropriate forms of storage" [Lavoie, 2004, p.8]), *data management* (the functional entity that maintains descriptive metadata regarding the AIP), and *preservation planning* (the functional entity that checks and ensures the preservation strategy or collection development policy is mapping to the AIP).

The archival information package is then transformed to a dissemination information package (DIP) via the *access* functional entity (which handles dissemination, information access, and requests from consumers). The DIP can interact with consumers (i.e., the designated community) directly. In addition, the *administration* functional entity oversees the day-to-day operation of all information packages (i.e., SIP, AIP, and DIP).

Table 2-3. Descriptions of the OAIS functional model

Micro	Meso		Macro
IP that OAIS interacts with	Functional entities	Function descriptions	External entities that OAIS interacts with (if applicable)
SIP, AIP	Ingest	the functional entity that accepts SIP submitted by the producers	Producer
AIP	Archival storage	the functional entity that ensures "archived content resides in appropriate forms of storage" (Lavoie, 2004, p.8)	--
AIP	Data management	the functional entity that maintains descriptive metadata regarding the AIP	--
AIP	Preservation planning	the functional entity that checks and ensures the preservation strategy or collection development policy is mapping to the AIP	Producer, consumer
AIP, DIP	Access	the functional entity that handles dissemination, information access, and requests from the consumers	Consumer
SIP, AIP, DIP	Administration	the functional entity that oversees the day-to-day operation of information packages	Management

Source: Data organized by this dissertation from Lavoie (2004)

Social science data repositories have adopted the OAIS model. As early as 2007, the world's largest social science data repository, the Interuniversity Consortium for Political and Social Research (ICPSR), published a series of articles and guidelines describing how ICPSR integrates the OAIS model into their work model. The outcome, the "ICPSR Pipeline," adopts the OAIS reference model in social science research data, and is well-documented in both "Designing the Future ICPSR Pipeline Process" (Gutmann, Evans, Mitchell, & Schürer, 2009) and "ICPSR meets OAIS" (Vardigan & Whiteman, 2007).

Aside from assessing compliance, there are existing studies that use OAIS as a foundational framework to examine data repositories' practices. For example, Yoon and Tibbo (2011) conducted a content analysis on the data submission package (SIP) elements, and examined submission forms and submission guidelines collected from 14 data repositories in the social science domain.

2.3.2 Metadata standards in social sciences

Metadata is the key to ensuring that research data can be well-discovered, accessed, used, preserved, and disseminated. This section reviews the most common metadata standards in social sciences.

Figure 2-3 illustrates and summarizes the most common metadata standards that major disciplinary fields adopt outside the social sciences. Large social science repositories such as ICPSR and the Dataverse Network adopt the Data Documentation Initiative (DDI) as a metadata standard (Borgman, 2015). As of 2016, there are 39 DDI adaptors around the world.

Despite the common adoption of DDI, a wide range of metadata standards are available in the diverse social science subject areas. Figure 2-4 illustrates the relationship between these subject areas and applicable metadata standards. Each blue circle represents a metadata standard and each green circle represents a subject area in social sciences. Each edge indicates a subject area that adopts a certain type of metadata. The data used to generate this diagram is obtained from the Digital Curation Centre (DCC)'s metadata standards, *Social Science & Humanities* (DCC, 2017).

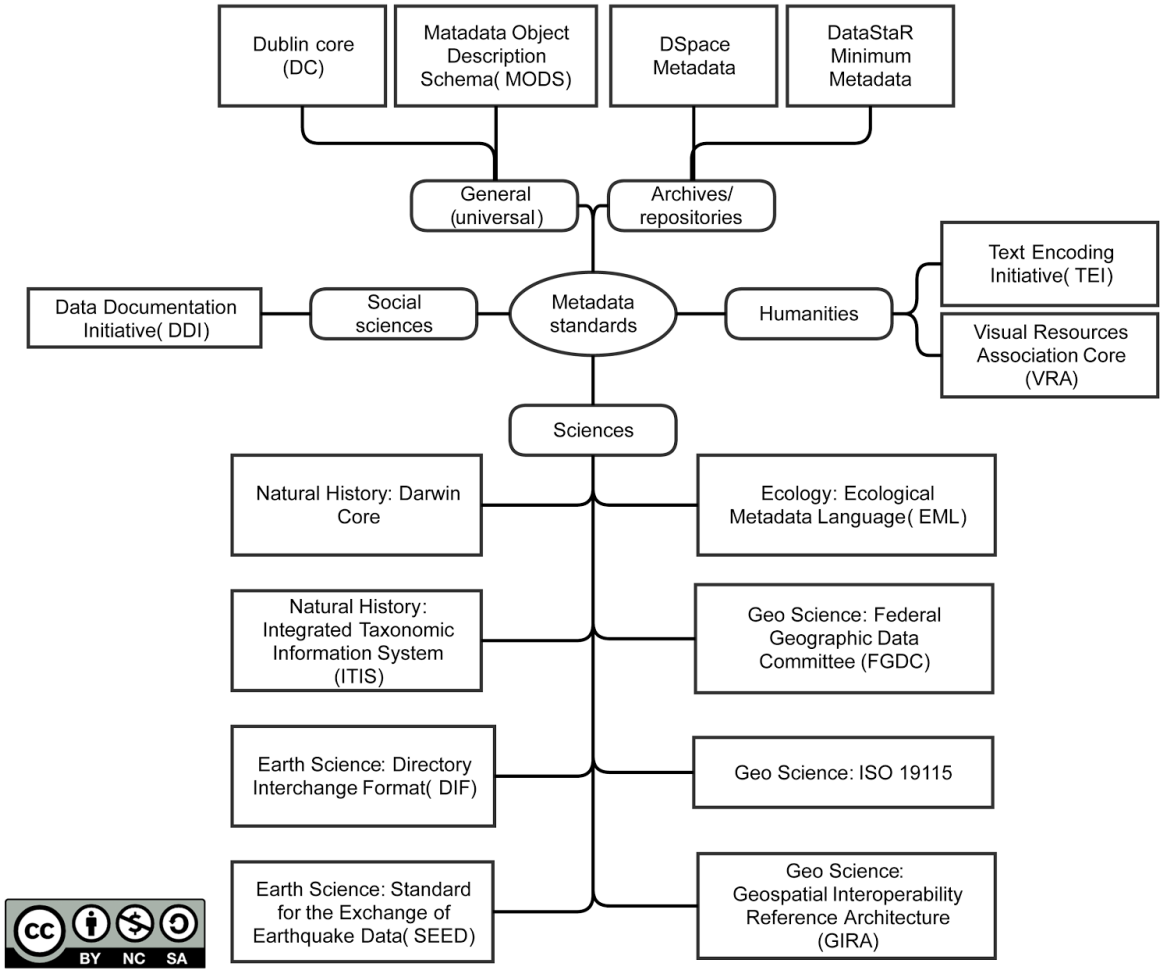


Figure 2-3. Most common disciplinary data standards

Source: Drawn by this dissertation, based on Metadata Concept Map by Amanda Tarbet, also under ShareAlike 3.0 License.)

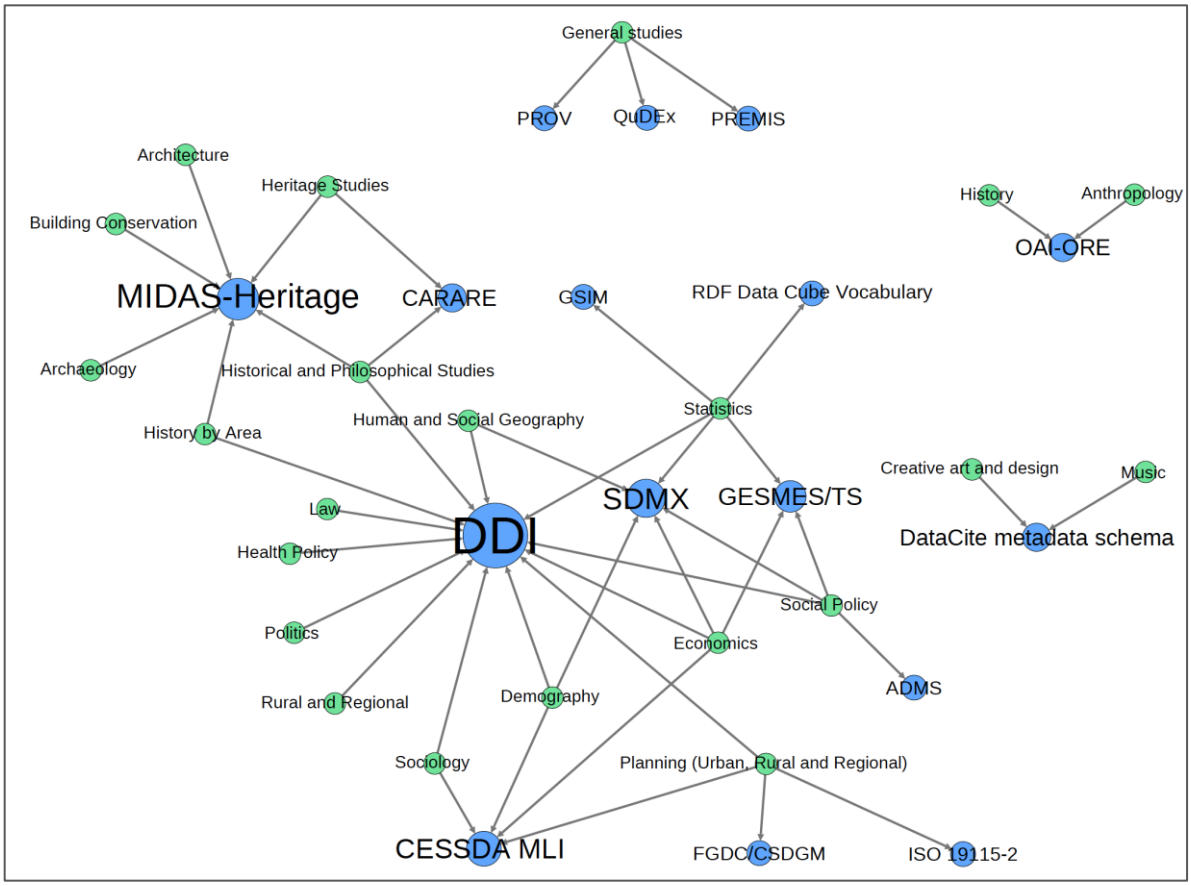


Figure 2-4. Social science metadata standards: fields and metadata

Source: Data visualized by this dissertation, data were collected from DCC, 2017.

Based on the data and the visualization, DDI is indeed a popular standard that has been adopted by most subjects, such as law, political studies, and health policy. However, it is worth noting that several standards are popular among different subject clusters. For example, MIDA-Heritage is supported by architecture, archaeology, historical and heritage; CESSDA is supported by urban planning and sociology; Statistical Data and Metadata eXchange (SDMX) and GEneric Statistical MESsage for Time Series (GESMES/TS) are favored by quantitative-oriented subjects such as economics and statistics; DataCite metadata is chosen by music and art design.

In summary, even though data sharing in social sciences may occur less often than in STEM disciplines, its infrastructure and metadata standards are reasonably mature (Borgman, 2015). Moreover, several professional associations (e.g., DCC and Metadata Standards Directory Working Group) strive to manage and advocate for metadata standards.

2.4 QUALITATIVE DATA SHARING

Research norms are an important factor regarding data sharing because they shape scholars' behaviors and day-to-day decisions. Hence, reviewing literature on research norms illuminates why social scientists might hesitate to share data, particularly data that contain sensitive personal information. On the other hand, the growing recognition of "integrity through transparency" encourages data openness in all disciplines. The interplay of these two conflicting ethical considerations (confidentiality and openness) complicates the data-sharing practices in social sciences.

2.4.1 Qualitative research and data

There is no easy dichotomy between qualitative and quantitative data, since qualitative data can also involve information with numerical values (Richards, 2014). Qualitative research usually addresses factors unlike those addressed by quantitative research. Instead of focusing on quantities, weights of factors, causes, and strengths of relationships, qualitative research explores issues, provides an understanding of phenomena, and answers questions by analyzing and interpreting unstructured information (e.g., information that is text-heavy or not organized in a pre-defined variable list) (Barbour, 2007). This information, regarding the whys and hows of human behavior, opinion, and experience, are usually difficult to gather through quantitative-oriented approaches (Guest, Namey, and Mitchell, 2012).

The difference in worldview between qualitative and quantitative research is not the only distinction between the two; the data generated from the former also has its unique types. In Ryan and Bernard's (2000) work, qualitative data are divided into three types based on format: audio, text, and video. Additionally, text analysis was subdivided into primary elements: text as proxy for experience (e.g., structured interviews) and text as object of analysis (e.g., narratives or online content).

Patton (2001) has a similar taxonomy to describe qualitative data, suggesting that it includes three kinds of formats:

- In-depth responses
- Direct observations
- Documents

In Patton's view, in-depth responses refer to open-ended questions that yield detailed feedback about people's opinions, experiences, feelings, and perceptions, whereas direct

observations come from researchers' fieldwork descriptions of people's activities, behaviors, and conversations. Documents, the third format, are written materials such as reports, publications, or human records e.g., clinical records or trial records.

While Patton describes three categories of qualitative data depending on how the data are generated, Holliday (2007) presents a list of raw data types to discuss "what counts as (qualitative) data" (p.60). His taxonomy, compared with Patton's, captures the original context and the actors. For example, the list includes five types of "researchers' description" regarding 1) descriptions of people's behavior, 2) descriptions of an event, 3) descriptions of institution (e.g., how a school is operating), 4) descriptions of appearance (e.g., number of green plants in an office facility), and 5) descriptions of research events (e.g., researchers' observations during a focus group). The remaining categories in Holliday's taxonomy also include people's actual words (e.g., responses on questionnaires or participants' diaries), audio records, visual records, and documents.

In *A Guide to Sharing Qualitative Data* at the Qualitative Data Repository (QDR), Elman and Kapiszewski (2013) use more concrete examples to enumerate common qualitative data. While not exhaustive, types of qualitative data include: "data from structured, semi-structured, or unstructured interviews such as audio, images, video, and text; focus groups; oral histories", "field notes (including from participant observation or ethnography)" (p.1) and other types of textual information or unstructured pieces. QDR further enumerates around 30 kinds of qualitative data that can be archived and shared (see Appendix A).

Many repositories like QDR provide concrete examples of qualitative data. The ICPSR suggests nine overarching types of qualitative data that are suitable to archive for reuse (ICPSR, n.d.):

- In-depth/unstructured interviews, including video

- Semi-structured interviews
- Structured interview questionnaires containing substantial open comments
- Focus groups
- Unstructured or semi-structured diaries
- Observation field notes/technical fieldwork notes
- Case study notes
- Minutes of meetings
- Press clippings

After collecting three types of literature sources (namely: discipline data repositories, academic articles, and funding agencies), Table 2-4 compares the definition of data from these three sectors.

Table 2-4. Types of qualitative data in this dissertation study

Types of qualitative data	Some Examples	Discipline Data Repositories: What should be archived?		Researchers: What are qualitative data?			Funding agencies: What counts data?		
		QDR	ICPSR	Hollidays	Patton	Ryan & Bernard	NSF-SBE	IES	NEH
Researchers' description	Observation notes Field notes Researchers' recordings Case study notes	◎	◎	◎	◎	◎	◎	◎	◎
Objects representing participants' view (Participants' actual words)	Transcription from an interview Interview recording (audio) Sketches or drawing Open-ended responses Participants' diaries Letters	◎	◎	◎	◎	◎	◎	◎	◎
Data related to methodology and data processing	Researchers' coding schemes Instruments Interview or focus group protocol	◎	◎			◎	◎		◎
Data related to documentation of the course of research	Minutes of meetings Internal memos	◎	◎						
Documents and real world objects	Court records Press clippings Clinical records Military records Maps Photographs	◎	◎	◎	◎	◎			

To sum up, previous work helps clarify the scope of qualitative data. When this dissertation refers to the term *qualitative data* in social sciences, it indicates data generated from both qualitative and mixed-method studies. While not exhaustive, common qualitative data types include:

- Data generated from researchers' descriptions (e.g., observation notes, field notes)
- Objects representing participants' views (e.g., participants' actual words, such as transcription from an interview or open-ended responses)

- Data related to methodology and data processing (e.g., researchers' coding schemes or instruments)
- Data related to documentation of the course of research, and other types of textual information or unstructured pieces
- Documents (e.g., court records, press clippings or clinical records)

2.4.2 Debates of survey questionnaire: quantitative or qualitative

While qualitative data are usually defined by enumerating examples, it is not always apparent whether a survey questionnaire study is qualitative or quantitative. Following the notions of qualitative data that Holliday and Patton describe, a questionnaire counts as qualitative data because it provides in-depth responses or reflects people's actual words. However, Elman and Kapiszewski (2013) do not mention survey questionnaires as a type of qualitative data in QDR. Jansen (2010) points out this potential confusion and provides contexts in which a survey can be classified as either quantitative or qualitative:

- Survey questionnaires are quantitative studies when they "primarily aim at describing numerical distributions of variables" (de Vaus, 2002, pp.3-7, as cited in Jansen, 2010). Statistics of the samples and "precision of estimates" (confidence level) are the main components in such a category (de Vaus, 2002, as cited in Jansen, 2010).
- Survey questionnaires are a qualitative approach when they determine "the diversity of some topic of interest within a given population" (Groves et al. 2004, p.3, as cited in Jansen, 2010). This type of survey does not aim at presenting a "number," but establishes "the

meaningful variation (relevant dimensions and values) within that population" (Groves et al. 2004, p.3, as cited in Jansen, 2010).

In addition, many survey questionnaires contain open-ended questions that allow participants to express in-depth opinions; thus, there is no clear-cut distinction between quantitative and qualitative classifications for survey questionnaires. Therefore, to be more comprehensive, this dissertation study includes survey questionnaires in the discussion of qualitative research.

To conclude, even though there are working definitions for research data, most U.S. funding agencies agree on the federal government's definition: data are evidence for the research community to validate research findings or to reuse for new studies. In addition, given that qualitative data has very different characteristics than its quantitative counterpart, developing universal guidelines to encourage data sharing might not reflect the different and difficult nature of qualitative data in social science disciplines.

2.4.3 The benefits of qualitative data sharing

Prior work has recognized the importance of qualitative data reuse and sharing (Elman & Kapiszewski, 2013). Reusing qualitative data provides scholars with an opportunity to reinterpret and "study the raw materials of the recent or more distant past to gain insights for both methodological and substantive purposes" (UK Data Services, n.d.). Besides reusing qualitative data, researchers and practitioners have identified several benefits of qualitative data sharing, including:

- Reanalysis, i.e., asking new questions or generating new findings built on the work of others, or by approaching the original data in "ways that were not originally addressed, such as using data for investigating different themes or topics of study" (ICPSR, 2010);

- Comparative research and meta-analysis (Bishop, 2009; Fry, Lockyer, Oppenheim, Houghton, & Rasmussen, 2009);
- Re-description or re-interpretation, i.e., describing the data again while considering "contemporary and historical attributes, attitudes and behavior of individuals, societies, groups or organizations" (ICPSR, 2010);
- Restudy of original research (i.e., on the same research question) to compare "with other data sources or providing comparison over time or between social groups or regions, etc." (ICPSR, 2010; Bishop, 2009; Fry et al., 2009);
- To ensure transparency and integrity of research procedures (Borgman, 2007);
- Verification on studies (Tsai et al., 2016);
- Methodological replication or advancement, e.g., following and replicating a focus group's protocol, designing innovative research approaches, reflecting, or enriching conventional methodology or research tools (ICPSR, 2010);
- Teaching and learning purposes, i.e., providing unique case studies or research materials for teaching and learning how to conduct research (Bishop, 2009; ICPSR, 2010).

However, despite the growth of qualitative data archiving and sharing (Rasmussen, 2011), today's researchers have still shown persistent skepticism about qualitative data archiving, reuse and sharing (Mason, 2007; Yoon, Hall, & Hill, 2014; Slavnic, 2011; Mauthner & Parry, 2009). Literature has identified and discussed several barriers and challenges that qualitative researchers encounter when archiving and sharing qualitative data.

2.4.4 The challenges of qualitative sharing

Scholarly debates about the barriers to archiving qualitative data have existed since the early 2000s. The arguments usually center on methodological challenges (i.e., the subjectivity in qualitative methodology) (Parry, Mauthner, 2004; Bishop, 2005; Parry & Mauthner, 2005; Mauthner & Parry, 2009), data ownership, informed consent, and confidentiality.

2.4.4.1 *Methodological challenges*

Qualitative scholars tend to rely on their own perspectives to understand, interpret, and explain the world. Instead of using an objective technique or a top-down strategy to examine the social world, Creswell (2009) explains that qualitative scholars usually apply a constructivist worldview, where "meanings are constructed by human beings as they engage with the world they are interpreting" (p.8). That is, no matter the identity of the researcher or the subject, humans make sense of behaviors and other phenomena based on their own perspectives.

Such subjectivity plays an important role and "guides everything from the choice of topic that one studies, to formulating hypotheses, to selecting methodologies, and interpreting data" (Ratner, 2002, para 1). This influences how qualitative researchers view and value their research data, and can result in resistance to qualitative data archiving and sharing, as explained below.

Qualitative data, the product of qualitative methodology, is also closely connected to the researcher or original research team (Fink, 2000). Broom, Cheshire, and Emmison (2009) conducted a focus group study to investigate qualitative scholars' practices for data archiving and sharing at Australian universities. The team found that qualitative researchers often see their own data as "organic," "intimate" and "personal," and expressed concern that others may misinterpret their data.

Furthermore, qualitative data consumers do encounter barriers while trying to reuse others' data. Yoon (2014) interviewed eight researchers who reuse qualitative data generated by others, and found that these researchers often heavily rely on the original sharers to gather sufficient understanding of the data. Yoon's study suggests there are challenges for a data consumer to reuse others' data if the data sharer fails to provide sufficient contextual information.

Using others' data might be also problematic. Research (e.g., Corti, 2000; Bishop, 2006) suggests that context can help recreate the original investigator's experience and can therefore compensate for an "absence" in the original research. Such contextual aids can be filed notes, photos, notes for research background, or annotations with interview transcripts (Corti et al., 2014). Regardless of the overhead caused by adding contextual information, other researchers argue that from an epistemological standpoint, context can never be replicated (e.g., Mauthner & Parry, 2009). Specifically, they contend that adding contextual background material as metadata does not successfully "overcome the epistemological problems of reusing data," nor does it allow for re-experience, since an investigator never directly engages with the original context (Mauthner, Parry & Backett-Milburne, 1998 as cited in Cheshire, 2009, p.31).

2.4.4.2 *Data ownership*

Another challenge of qualitative studies is the ambiguity of data ownership, which exists in two granularities.

The first level of ambiguity resides between the institutions and individual researchers. As Cliggett (2013) states, scholars doing quantitative studies tend to think that data belong to the institution, whereas qualitative scholars think data belong to individual researchers. The second level of ambiguity resides between the researchers and their human subjects. Since qualitative scholars work closely with their human subjects and the study results might be a joint endeavor or a "co-

production" between the two (Moore, 2007; as cited in Broom, Cheshire, & Emmison, 2009), it is unclear who – the researcher or the participant/informant – owns the data.

Parry (2004) raises this concern by using an interview as an example. The copyright of an interview recording can be separated into two parts: the spoken word and the production of the recording. Obviously, the informant who was interviewed and the interview mediator share the former copyright as joint speakers. However, the researcher, who conducted and designed the interview protocol, owns the copyright of the recording. Corti et al. (2014) suggest that qualitative researchers could resolve this early on by informing the participants about what will become of the data and gaining permission from them. For example, researchers could tell participants that the data will be archived or shared under certain identity protections.

However, for the studies that intentionally keep participants unaware, it might be difficult to get informed consent upfront. The next section discusses the debates and challenges regarding informed consent.

2.4.4.3 *Confidentiality and anonymity*

Ensuring confidentiality and anonymity (i.e., protecting the identity of study participants) are critical requirements for data archiving and sharing, the importance of which has been recognized by an increasing number of federal laws, such as the UK's Data Protection Act (Corti et al., 2014; Parry & Mauthner, 2004). Some data repositories also demand that personal information about research participants remain confidential, to guarantee that "confidentiality and anonymity are to be honored" (Parry & Mauthner, 2004, p. 143). However, Cliggett (2013) points out that protecting study participants, confidentiality, and data with sensitive information are the most frequently-mentioned barriers by qualitative researchers regarding data archiving and sharing.

There is still an inherent dilemma between hiding personal information (to achieve anonymity and confidentiality) and preserving contextual information (to ensure subjectivity), as some key characteristics of participants might be important for data consumers to understand context.

2.4.4.4 *Informed consent debates*

Israel and Hay (2006) reviewed several community guidelines, including the American Sociology Association and four other professional communities, concluding that "most guidelines for ethical research require all participants to agree to research before it commences" (p.61). Informed consent, according to Israel and Hay, implies two related activities:

- "Participants need first to comprehend the nature of the research" (p.61), and
- "Participants need second to agree voluntarily to the research" (p.61).

Obtaining genuine informed consent from the participants may be very legitimate, but many researchers in the humanities and social sciences find it difficult to inform participants before the observation begins. For example, some social experiments aim to test people's reactions or true attitudes in a natural situation, and "such consent has damaged their (social scientists') research and has not been the best interest of research participants" (Israel and Hay, 2006, p.60). Some of these studies must be carried out on unwitting subjects to fulfill the research goal.

Hence, the researchers of such studies are either unable to obtain initial informed consent or are forced to obtain consent that may contravene the actual research objective. To minimize ethical concerns, explanations are often given at the end of the study. Still, without informed consent, the ownership of the data that has been produced is ambiguous, and this ambiguity makes researchers reluctant to share their qualitative data. To reduce the ambiguity of data ownership and to fully inform participants, Corti et al. (2014) suggest that no matter how informed consent is obtained

(e.g., verbal or written), the researcher should notify the participants about any unknown future uses of the data.

2.4.5 Qualitative sharing data sharing at national and institutional levels

Data sharing can be broadly classified into two types, in terms of the channel choice: formal and informal. Formal sharing is the process of publishing data to data repositories or in academic journals as appendices. Informal sharing, however, often involves unofficial communication channels such as submissions to personal websites or sending data to others upon request.

Even though sharing qualitative data began later than its quantitative counterpart, qualitative data repositories have emerged in recent years. For example, the UK's ESRC Data Service was founded in the 1960s, but a dedicated qualitative data center, ESDS Qualidata (UK), was not established until the 1990s (Hammersley, 1997). ESDS Qualidata was later merged into the UK's ESRC Data Service in 2012 (Qualidata, 2012), indicating the recognition of qualitative data practices from governmental departments.

Like the UK, many countries—including Ireland, Austria, Finland, Australia, and the U.S. (summarized in Table 2-5)—have also gradually increased their investments in qualitative data repositories since the 2000s (Broom, Cheshire, & Emmison, 2009). Countries such as Denmark, Germany, Switzerland, and Norway are also on their way to establishing qualitative data repositories.

Table 2-5. Data repositories for archiving qualitative data

Counties	Data centers for archiving qualitative data	Year established/started acquired qualitative data*	Reference
U.K.	QualiData; UK Data Services	1990; 2012*	UK Data Services, n.d.
Ireland	Irish Qualitative Data Archive	2008	O'Carroll, 2011
Austria	Wiener Institute for Social Science Data Documentation and Methods (WISDOM)	2008	Smioski, 2011a; 2011b
Finland	Finnish Social Science Data Archive	2003*	Kuula, 2011
Australia	Australian Data Archive-Qualitative	2010	ADA, n.d.
U.S.	Inter-university Consortium for Political and Social Research (ICPSR)	2011	ICPSR, n.d.
U.S.	Qualitative Data Repository (QDR) at Syracuse University	2013	QDR, n.d.

Besides national data archives or disciplinary data archives supported by governmental agencies such as NSF, there are also data archives supported by disciplinary communities. For example, the International Association for Social Science Information Services and Technology (IASSIST) has a mission to advance "information technology and data services to support research and teaching in the social sciences" (IASSIST, n.d.). Also, a professional group called the DDI Qualitative Data Model Working Group helps formulate a "robust XML-based schema for qualitative data exchange" (QDMWG, n.d.).

However, archiving qualitative data in a formal manner is not a popular practice among social scientists (Broom, Cheshire, & Emmison, 2009). This observation is consistent with other scholars who argue that it is an uncommon practice to formally share qualitative data, such as formally publishing data or sending data to archives. Kjeldgaard (2010) reviewed qualitative data-sharing practices in Denmark and concluded that "neither qualitative data sharing nor reuse is practiced formally" (p.39). However, there is a lack of literature that specifically studies how qualitative scholars informally share their data.

Despite increasing attention on qualitative data repositories and qualitative data-sharing studies, *qualitative data reuse* remains rare due to the barriers discussed in Section 2.2.2. Curty, Kim, and Qin (2013) conducted a mixed-method study involving a survey questionnaire for PIs and a content analysis on NSF awardees' data management plans (DMPs). They received 169 responses and analyzed 68 DMPs. Their survey results reveal there are many barriers to data reuse: anonymity, confidentiality, lack of context and documentation, extra time effort, lack of tools, and lack of interoperability and standards. These results highlight the vital need to conduct a comprehensive study on the entire ecosystem of qualitative data sharing, including archiving, sharing, and reuse.

In summary, current data centers and research still lack empirical studies on qualitative data sharing. This dissertation study aims to compensate this vacancy.

2.4.6 Best practices for qualitative data sharing

Despite the existence of guidelines or best practices for general data management at university libraries, there is still a lack of guidelines designed for qualitative data sharing (Slavnic, 2013; Yoon et al., 2014). A few examples exist and most adopt or cite QDR's "A Guide to Sharing Qualitative Data." However, this guideline is customized to this particular repository.

The QDR qualitative data sharing guideline dedicates space to instruct PIs on how to deal with ethical issues (e.g., adding data-sharing into the consent process and anonymization) and data ownership. For example, QDR encourages potential PIs to obtain consent before interviewing the participants, but for those who have already conducted research, QDR suggests that "scholars should determine to what degree the Institutional Review Board agreement and protocols associated with collecting those data would cover such sharing, and discuss the process for gaining permission retroactively with IRB staff" (p.5).

For the anonymization of qualitative data, QDR suggests that researchers replace direct identifiers with descriptive replacement terms or otherwise generalize the details. QDR provides the following examples: "replacing a doctor's detailed area of medical expertise with an area of medical specialty" and "creating an anonymization log (stored separately from the anonymized data files) of all replacements, aggregations, or removals" (p.7).

However, the QDR guideline focuses on the ethical issues and data ownership checks, and pays relatively less attention to the discussion of data characteristics, the preferred scope of qualitative data for the social science community, or what kinds of qualitative data are more useful than others. These are left up to the PIs' discretion.

While the QDR guideline is repository-centered, the library guide at the University of California, Berkeley (Cal), "Managing and Sharing Qualitative Research Data 101," covers "which qualitative data should I keep and share," and can be viewed as a research-centered guide. Since this guide helps potential PIs think about how unique their data are and how data users can gain insights from the data and draw similar conclusions, it might be more practical for researchers than the QDR guideline.

The UK archive's "Sharing Qualitative Data Challenges and Opportunities" (Bishop, 2016) mentions qualitative data sharing "in accordance with relevant standards and community best practice" (p.10). Yet, these standards and best practices do not exist, creating a circular dependency (i.e., data management guides refer users to disciplines for the best practices, but the discipline looks for solutions in such data management guides). Nevertheless, Bishop (2016) still advises how to deal with confidential data by checking four components before depositing qualitative data: obtain informed consent, protect identities, regulate access, and securely store. Most importantly, one "should not place unreasonable burden on primary researchers" and ensure that "funding is available" (p.20).

In addition to the considerations of repositories and researchers, some work discusses data sharing from the journal venue's perspective. Tsai and his colleagues (2016) in their article "Promises and pitfalls of data sharing in qualitative research" suggest that journal editors develop a minimum requirement in a journal's data-sharing policy for qualitative research. The discussion study is in the context of bio-medicine, which has a similar need to establish best practices for qualitative data sharing. Tsai et al.'s recommendations can be summarized below (p.196):

1. Require authors to provide a statement explaining whether consent to share data was obtained from the participants.
2. Require authors to carry out "minimum standards for deidentification" and "[e]ncourage anonymization of field notes."
3. Encourage authors to recruit multiple informants and/or informants from different institutions (i.e., hospitals) to reduce the risk of direct identification.
4. Allow authors to share their coding results "as an alternative to full (interview) transcripts."
5. "Encourage authors to document social audits or other stakeholder dissemination."
6. Ensure manuscript reviewers with expertise or experience "in qualitative and mixed methods research to comment on the adequacy of anonymization."
7. Establish a petitioning process for non-disclosure of data.

In summary, though there are several guidelines from the perspective of repositories, researchers, or journal venues, they are mostly only focusing on the legal and ethical aspects. Individual disciplines must continue the discussion about the value of data and data ownership to arrive at a consensus for developing a better practice in qualitative data sharing.

2.5 IMPLICATIONS FOR RELATED WORK

Several implications can be drawn from the above literature review. First, most prior work presents authors' research notes rather than empirical studies (e.g., Hammersley, 1997; Parry & Mauthner, 2004; 2005; Mauthner & Parry, 2009; Bishop, 2005, 2007, 2009; Heidorn, 2008). While these opinion-based or reflection papers provide valuable viewpoints on qualitative data sharing, empirical surveys are still needed to gather feedback from field researchers and to understand actual data-sharing practices.

Second, a large fraction of existing empirical studies focus on STEM disciplines (e.g., Tenopir et al., 2011; Tenopir et al., 2015; Kim, 2013; Sayogo & Pardo, 2012; Wallis, Rolando, & Borgman, 2013) and thus, conclusions drawn from these studies might not be applicable to social science disciplines.

In addition, using mixed methods is legitimate to investigate scholars' data-sharing practices. Prior work has successfully used various research methods, ranging from Web content analysis, questionnaires, and in-person interviews to observation methods or ethnography. Hence, it is anticipated that combining several research methods (referred to as mixed methods hereafter) is feasible and can lead to comprehensive results.

In summary, there is an imperative need to bridge the gaps between research on quantitative and qualitative data sharing, and between STEM and social science disciplines. After reviewing the related work on data sharing, the inadequate literature in the realms of qualitative data, social sciences, and humanities becomes evident.

3.0 LITERATURE REVIEW II:

CONCEPTUAL FRAMEWORK FOUNDATIONS

The second part of the literature review presents the conceptual framework foundations for this dissertation study, including framework to support data sharing in the digital environment, profiling tools for data practice, and motivation theories. The goal of this chapter is to identify the theoretical framework, which provides a normative framework to guide the research design.

This dissertation study requires a theoretical framework for scholarly collaboration in digital environments with the following two properties:

- This framework should help identify key dimensions which can then construct potential factors and actual questions.
- This framework should be applicable to diverse social research methods; for example, it can be used for a survey, interview, focus group, or content analysis.

Because most previous work focuses on numerical data or STEM disciplines, the first step in this dissertation study is to build a preliminary framework tailored for qualitative data. Two theories—Knowledge Infrastructures (KI) including seven elements and Olson's Theory of Remote Scientific Collaboration (TORSC)—serve as a high-level abstraction guiding the framework design (Section 3.1). Section 3.2 discusses the data practice profiling tools which help identify critical factors that need to be covered. DCP and CCMF are used to develop items under this framework, the advantage of which is that items in CCMF focus on technological and organizational infrastructure and sufficiently cover legal and funding aspects. The components in DCP are then

used to collect data characteristics. However, both are lacking considerations about individual motivations, and therefore motivation theories are introduced to fill this vacancy. Motivation theories in Section 3.3 provide a theoretical foundation for constructing specific items related to individual scholars' motivations.

3.1 FRAMEWORK TO SUPPORT DATA SHARING IN DIGITAL ENVIRONMENT

3.1.1 Knowledge Infrastructure (KI)

The term "knowledge infrastructure" builds on the earlier development in e-Research movements and information infrastructure (Borgman, 2015). Transformed from information infrastructure (Bowker, Baker, Millerand, & Ribes, 2010), knowledge infrastructures refer to "robust networks of people artifacts and institution that generate, share, and maintain knowledge about human and natural worlds" (Edwards, 2010, p. 17, as cited in Borgman, 2015). Knowledge infrastructures include seven elements – people (individuals), shared norms and values, artifacts, institutions (organizations), routines and practices, policies, and built technologies – all of which work together as a complex ecology (Edwards et al., 2013; Borgman et al., 2014).

Scholars use KI to make sense of knowledge-sharing mechanisms. Ribes and Finholt (2009) used KI to evaluate how projects are run and how knowledge is preserved. They conducted a series of case studies on four national research projects on infrastructure development in the digital environment: GEON (Geosciences Network), LEAD (Linked Environments for Atmospheric Discovery), WATERS (Water and Environmental Research Systems), and LTER (LongTerm Ecological Research). Following the KI framework, the authors examined and compared these four

projects based on their facility, community interests, technological readiness, and production quality systems.

Others consider KI to be a holistic framework to interpret how current technology can support their researchers. Australian researchers Wolski and Richardson (2014) discuss how the related components in KI, such as organizational structure, built infrastructure, digital artifacts, and people, can fit into new forms of digital scholarship. From their perspectives and insights on built technologies, infrastructures such as "Infrastructure as a Service (IaaS), Software as a Service (SaaS) or Platform as a Service (PaaS)" are sorely needed by scholars who work in the digital environment. These techniques provide a flexible, tailored, and accessible infrastructure for individual scholars or institutions, a concrete example of which is cloud storage.

3.1.2 Theory of Remote Scientific Collaboration (TORSC)

Data sharing can be considered a kind of scholarly collaboration. This dissertation study adopts the Theory of Remote Scientific Collaboration (TORSC) to enrich and complete the theoretical foundation of KI by considering more elements of scientific collaboration.

Olson and Olson (2000) discuss four concepts that lead to success in remote scientific collaboration:

- common ground,
- coupling work,
- collaborative readiness, and
- technological readiness.

These four concepts have been adopted by the fields of information science and behavioral science by researchers who want to discuss the essence of scholarly collaboration and

communication (Borgman, 2007). Later, in 2008, Olson and his research team came up with TORSC, which extends the research context in their 2000 framework to general collaboratories. The updated framework comprises five overarching categories: the nature of work, common ground, collaboration readiness, management/planning/decision making, and technological readiness (Olson, Zimmerman, & Bos, 2008, p.80).

The "nature of work" category evaluates whether the work is unambiguous, whether collaborators can work independently, and whether it is a tightly-coupled work or not. The "common ground" category evaluates whether collaborators share common vocabularies and working patterns or management style. The "collaboration readiness" category evaluates whether there is a common goal, whether participants find each other to be reliable to work with, and whether there are motivations for collaborators. Olson and Olson (2000) argue that "different fields and work settings engender a willingness to share" (p.164). The "management/planning/decision making" category evaluates leadership, whether the distributed collaborators can communicate often, and whether there is an easy-to-reach contact channel. The "technological readiness" category evaluates whether the collaboration technologies provide functionality and ease of use, and whether participants are comfortable using them.

In this dissertation study, these five elements of TORSC can help create a framework to support digital scholarship. A more detailed application of the study is introduced in Chapter 5.

3.2 PROFILING TOOLS FOR CAPTURING RESEARCH DATA PRACTICES

3.2.1 Community Capability Model Framework (CCMF)

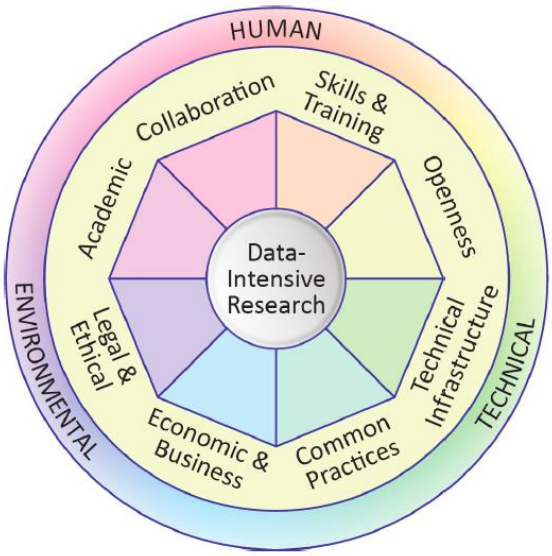


Figure 3-1. Community Capability Model Framework (CCMF)

Source: Jeng & Lyon, 2016

The Community Capability Model Framework (CCMF), developed by UKOLN Informatics and Microsoft Research Outreach (previously known as Microsoft Research Connections), aims to examine the infrastructure of an academic discipline's data curation, management, and sharing practices (Lyon, Ball, Duke, & Day, 2012).

The framework discusses eight relevant factors for determining the capability or readiness of a community to perform data-intensive research (see Figure 3-1):

1) Collaboration, in which participants describe their collaborative cultures between sectors, themselves and their colleagues, and if their studies engage the public.
2) Skill and training, in which participants are asked to assess their own skill sets and evaluate their institutional training programs related to data curation.
3) Openness, in which participants are asked to provide the extent of openness regarding their research, methods, data, and research outcomes.
4) Technological infrastructure, in which participants are asked to evaluate their discipline-wide support in terms of data storage, computing, processing, discovering, and accessing.
5) Common practices (in data management), which captures participants' data characteristics and how they describe their data.
6) Economic and business models, in which participants are asked to answer questions related to funding, in terms of scale, location, and coverage.
7) Legal, ethical and commercial factors, in which participants answer questions related to regulatory framework, norms, and ethical responsibilities.
8) Research culture, in which participants are asked to answer questions related to reward models and entrepreneurship.

The CCMF Toolkit was released as an instrument, in a spreadsheet style, that includes a consent form, 10 open-ended questions about an interviewee's data profiles, and 55 other questions related to the eight critical factors. In the applications of this toolkit, Brandt applied CCMF to study agronomy scholars' data practices and eight capability factors. His findings have been presented at the Research Data Access & Preservation Summit 2014 (as cited in Lyon, Patel, & Takeda, 2014).

Not only can CCMF be an equipped instrument that examines technological readiness for data practices in a discipline, but it can also be a comprehensive framework that helps researchers assess and understand a discipline's or institution's capacity for supporting data-intensive research.

3.2.2 Data Curation Profiles (DCP)

Data Curation Profiles (hereafter: DCP) is a tool that supports the assessment and analysis of data characteristics and scholars' discipline data practices (Cragin et al., 2010; Witt, Carlson, Brandt, & Cragin, 2009). Another apparent use for each completed profile is as a resource helping other professionals quickly capture how specific data are generated and used/reused in a certain research discipline. The description on the DCP website (http://datacurationprofiles.org) states:

> *"A Data Curation Profile is a resource for Library and Information Science professionals, Archivists, IT professionals, Data Managers, and others who want information about the specific data generated and used in research areas and sub-disciplines that may be published, shared, and preserved for reuse."*

In the DCP Toolkit, the persons interviewed and whose insights and data practices are represented in the data curation profile are called "data clients." Persons who interview, transcribe data clients' information and complete the profile are called "DCP researchers." The typical form of a DCP comprises the following elements:

- Section 1: Summary of data curation needs, which the data clients are asked to provide.
- Section 2: Overview of the research, in which DCP researchers ask clients to describe their research area focus, intended data audiences, and funding sources.
- Section 3: Data kinds and stages, in which researchers ask data clients to explain their data characteristics, including their research process and "the context of how the data is used in the data client's research" (Data Curation Profile, n.d., p.5).

- Section 4: Intellectual property context and information, in which data clients describe data ownership and conditions for access and reuse.
- Section 5: Organization and description of data, in which data clients provide an introduction about how their data are described and organized, and if there are formal standards used.
- Section 6: Ingest/Transfer, in which researchers ask data clients to share any "preparations or actions needed before the ingestion or transfer of data would take place" (p.9).
- Section 7: Sharing & Access, in which researchers capture data clients' willingness and motivation to share data.
- Section 8: Discovery, in which data clients narrate their general need and approaches for data discovery.
- Section 9: Tools, in which data clients describe how they generate data, focusing on the tools that help them collect, process and analyze data.
- Section 10: Linking/Interoperability, to determine if their data are linking or interoperating with other datasets.
- Section 11: Measuring Impact, in which researchers can determine if data clients have any needs or preferences regarding the metrics that measure their data's impact.
- Section 12: Data Management, in which data clients narrate how the current data has been managed, including security and back-up details.
- Section 13: Preservation, in which data clients describe if there is a need for long-term presentation of the data.

So far, the DCP website has published five volumes since 2009 and includes seven profiles related to social sciences and humanities[1]. Researchers Lage, Losoff, and Maness (2011), in the University Libraries at the University of Colorado-Boulder, have also adopted the DCP tool to exanimate the institution's scientific data curation activities. Their findings, presented as eight persona profiles, help academic librarians and data librarians understand data clients' needs, barriers, and data-related activities.

As with the CCMF Toolkit in Section 3.2.1, the DCP tool also functions to evaluate data curation practices from the perspective of the producers. Such profiling tools can help this dissertation study in developing concrete and comprehensive measures for a holistic investigation of scholars' data sharing.

3.3 MOTIVATION THEORIES

Human behavior is complex, and the cause of certain behaviors interests many researchers who study motivations. According to *The SAGE Glossary of the Social and Behavioral Sciences*, "theoretical models of motivation attempt to explain why individuals choose to engage in a particular activity" (p.333). Previous effort from researchers has shown that external events, personalities, job characteristics, and prior life experiences can affect people's motivations and behaviors (Kanfer, Chen, & Pritchard, 2008; Latham & Pinder, 2005).

[1] Zilinski & Lorenz (2012): Linguistics; Sapp Nelson & Beavis (2013): History; Eaker (2012): Architectural History; Jenkins (2012): Sociology; Tancheva (2012): Linguistics

As Borgman (2015) states, scholars find it difficult to justify data sharing as a return of investment. To understand their decisions about sharing data, it is crucial to investigate individual concerns, perceived efforts, attitudes, and expectations regarding data sharing. In this section, two well-known motivation theories are discussed and used to examine individual social science scholars' data-sharing intentions and motivations.

3.3.1 Intrinsic and extrinsic motivations

Human motivations can be categorized into three types, according to the Self-Determination Theory (hereafter: SDT): amotivation (i.e., without motivation), intrinsic, and extrinsic (Deci & Ryan, 1985; Ryan & Deci, 2000). This review focuses on the concepts of intrinsic and extrinsic motivation, which have been studied intensively over the past decades (Ryan & Deci, 2000). Prior studies have revealed the differences between intrinsic and extrinsic motivations, and have shed light on organizational knowledge sharing.

While intrinsic motivation originates from one's interests (e.g., for fun), psychological needs (e.g., inherent satisfaction or sense of belonging), and personal curiosities, extrinsic motivation "arises from environmental incentives (e.g., rewards) and consequences (e.g., reputations)" (Reeve, 2005, p.134). As Reeve (2005) further defines, incentives do not directly cause behaviors; they might increase the likelihood of whether a response will be triggered or initialized.

This dissertation study adopts this distinction between motivations for two reasons. First, the SDT has undergone public scrutiny since its conception and has been applied to the field of knowledge sharing. Second, this study can benefit from the SDT's distinction between intrinsic and extrinsic motivation: it can help distinguish the most critical motives that drive a social scientist (such as the mentioned research norms and benefits) when one investigates data-sharing behaviors.

The SDT framework can be applied to examine whether researchers are motivated extrinsically by an increase in citation counts, or simply by the sense of achievement from sharing great research (Elaman, 2010).

Lin (2007) examined employees' knowledge-sharing attitudes and intentions by using four factors: institutional rewards and expected reciprocal benefits (extrinsic factors), and self-efficacy and enjoyment of helping (intrinsic factors). Lin found that intrinsic factors are more effective than extrinsic factors in terms of knowledge sharing, and in fact, the expectation of reciprocal benefits has no association with knowledge-sharing attitudes and intentions. Further investigation is required to determine whether the same is true for why social scientists share data.

The Theory of Planned Behavior (TPB) is also raised frequently in the context of data sharing and knowledge sharing. For example, Gagné (2009) presented a conceptual model of knowledge-sharing motivations, which combines the SDT with the TPB.

3.3.2 Theory of Planned Behavior (TPB)

The Theory of Planned Behavior (TPB) concludes that "attitudes toward the behavior, subjective norms with respect to the behavior, and perceived control over the behavior are usually found to predict behavioral intentions with a high degree of accuracy" (Ajzen, 1991, p. 206). Behavioral intentions can further predict actual behavior.

In Ajzen's TPB (1991), the first determinant of behavioral intention is people's attitudes about the behavior (see Figure 1 in Ajzen, 1991). This refers to the extent to "which a person has a favorable or unfavorable evaluation or appraisal of the behavior in question" (p. 188). Another conceptual factor is subjective norms regarding the behavior, including normative beliefs, which refer to "perceived social pressure to perform or not to perform the behavior" (p.188). The third

determinant of behavioral intention is perceived control over the behavior. This can be understood as a predictor referring to people's perception of the "ease or difficulty of performing the behavior" (p.188).

Ajzen's TPB provides a conceptual framework for many researchers interested in scientists' motivations to share data. For example, de Montalvo (2003) adopted the TPB as a research framework to develop a model of spatial data sharing, which helped to "map out the belief structures underlying intentional behavior" (p.21). De Montalvo then customized the original TPB and identified three factors: 1) attitudes toward spatial data sharing, 2) social pressure from the research community, and 3) perceived control over spatial data-sharing behaviors. The result suggests that the TPB has been sufficiently applied into such a research context, and the customized model is also effective and generalizable, even for disciplines outside the GIS community (de Montalvo, 2003).

In subsequent research, Kim and Stanton (2012) conducted a mixed-method study (including interviews and a large-scale survey) to examine critical factors influencing STEM researchers' data-sharing practices. They specify two overarching themes (institutional factors and individual factors) to model scholars' willingness to share data. In terms of the individual, Kim and Stanton also adopted the TPB and customized three determinants as perceived benefit, perceived cost, and perceived risk: "Each of the determinants of behavioral intention is in turn influenced by underlying belief structures" (Kim & Stanton, 2012, p.48). They found that some researchers believe data sharing can highlight the quality of their research work. In contrast, researchers also believe that data sharing imposes a cost. Additionally, certain perceived risks prevented researchers from sharing their data with other researchers. Sayogo and Pardon (2013) also used TPB to explore challenges in terms of scholars' data-publishing behaviors. They obtained some interesting findings, including the lack of attention to proper acknowledgement and appreciation, since "researchers do not consider

acknowledgement and appreciation as an important determinant for publishing their research data online" (p.S26).

3.4 COMBINING FRAMEWORKS TO STUDY DATA SHARING

Prior work that investigates social science researchers' data sharing is missing a consolidated theory; thus, this dissertation study aims to compile a comprehensive study from diverse theories and tools. While some well-conducted studies have converged the TPB and the institution theory to explain individual data-sharing behaviors (e.g., Kim, 2013; Sayogo, 2012), theories behind the holistic model of data-sharing practices are still being explored and a consensus has not yet been reached. Similarly, data management profiling tools (i.e., CCMF and DCP) have advantages and concentrations. Combining these research tools is necessary for this dissertation study.

Inspired by prior research and the review of the theoretical foundations of KI and TORSC, this study propositions a four-dimensional framework that categorizes factors of data-sharing practices. The framework in Table 3-1 is used to investigate social scientists' data-sharing practices, including individual motivations and concerns, data characteristics, organizational contexts (specializing in discipline communities), and technological supports.

Table 3-1. Dimensions to study data-sharing practices

Applying to dimensions to studying data-sharing practices	Framework to support digital scholarship	
	Knowledge Infrastructure (KI)	Theory of Remote Scientific Collaboration (TORSC)
Individual motivations and concerns	Collaboration readiness	People (individuals) Shared norms and value
Data characteristics	Nature of the work	Artifacts
Organizational context (specializing in discipline community)	Common ground Management, planning, and decision making	Institutions (organizations) Routines and practices Policies
Technological supports	Technological readiness	Built technologies (system and networks)

4.0 PRELIMINARY STUDIES

4.1 OVERVIEW

This chapter describes two preliminary studies that shed light on the design of the main study.

The first preliminary study (PS1: Community Capability Study) examines the capability of scholars' communities and institutional infrastructures in terms of data production, curation, and management. Thirteen social scientists were invited to complete a survey and interview between June 2014 and April 2015. These scholars were asked to self-assess whether their academic environment provides supportive infrastructure for data curation. This assessment includes eight aspects: collaboration, skills & training, openness, technological infrastructure, common practices, economic & business, legal & ethical and research culture. The participants reported that their institutions have made relatively slow progress on economic support and data science training courses, but acknowledged that they are well informed about and trained for participants' privacy protection. The result of PS1 confirms a prior observation from the literature: social scientists pay close attention to ethical concerns, but lack technical training and support.

Another preliminary study (PS2: Research Process Study) aims to advance the understanding of how H&SS scholars collect, process, and interact with data at each stage of the research process, thus opening the "black box" on how they conceptualize their research processes and the data in their research. The sketches produced in this RPS study provide insight on the design of this

dissertation, and also identify opportunities for an academic library or data service provider to support H&SS scholars' research activities.

4.2 PRELIMINARY STUDY 1: COMMUNITY CAPABILITY STUDY

4.2.1 Research design

A pilot qualitative case study was designed in accordance with the Community Capability Model Framework (CCMF) developed by UKOLN Informatics and Microsoft Research Outreach (previously known as Microsoft Research Connections) (Lyon, Ball, Duke, & Day, 2012), which aims to examine the infrastructure of an academic discipline's data curation, management, and sharing practices.

4.2.2 Instrument modifications

The instrument covers eight factors contributing to data management capability, which were assessed to gain an understanding of data infrastructure issues in social science disciplines (Table 4-1).

Table 4-1. Eight dimensions of the CCMF instrument

#	Dimension	Description
1	Collaboration	Researchers describe the collaborative cultures between sectors, between themselves and their colleagues, and if their studies engage the public.
2	Skill and training	Researchers are asked to assess their own skill sets and evaluate their institutional training programs related to data curation.
3	Openness	Researchers are asked to describe the extent of openness regarding their research, methods, data, and research outcomes.
4	Technical infrastructure	Researchers are asked to evaluate their discipline-wide support in data storage, computing, processing, discovering, and accessing.
5	Common practices	Researchers capture details about their data characteristics and how they describe their data.
6	Economic and business models	Researchers are asked to answer questions related to funding, in terms of scale, location, and coverage.
7	Legal, ethical and commercial	Researchers answer questions related to regulatory framework, norms, and ethical responsibilities.
8	Research culture	Researchers are asked to answer questions related to reward models and validation framework related to their research.

This preliminary study adopts the CCMF Toolkit with discipline-tailored modifications that are designed primarily to enhance comprehension. This was achieved by adding social-science-friendly descriptions, exemplars, or tools and providing explanations of technical terminologies. There were 37 modifications in total; some sample modifications are provided in Table 4-2. Five capability levels are used to describe the level of ability or activity within a dimension: 1) Nominal Activity, 2) Pockets of Activity, 3) Moderate Activity, 4) Widespread Activity, and 5) Complete Engagement. The score for a particular capability factor indicates the perceived position of that community from the viewpoint of the researcher. A full version of the customized CCMF instrument is provided in Appendix B.

Table 4-2. Modification examples to CCMF

Modification Categories	Examples of Original Versions	Examples of Modified Versions
Adding discipline-tailored exemplars and tools	4.2 Tool support for data capture and collection 5.5 Standard vocabularies, semantics, ontologies	4.2 Tool support for data capture and collection (e.g., Screencasting tools, digital audio recorder, Web content scripters, Qualtrics, SurveyMonkey) 5.5 Standard vocabularies, semantics, ontologies (e.g., LCSH, MeSH)
Providing explanations of technical terminologies	2.11 Data referencing and citation e.g. DataCite DOIs 2.12 Data metrics and impact e.g. impact factors, altmetrics	2.11 Data referencing and data citation e.g. it uniquely identifies an object stored in a repository, such as DataCite DOIs) 2.12 The concepts of measuring scholarly impacts on data e.g. Impact factors of research datasets, altmetrics of datasets such as the number of downloads
Providing discipline-tailored descriptions in social sciences	3.4 Openness of methodologies/workflows (e.g short "how-tos", scripts for processing, programs for conversions)	3.4 Openness of methodologies/workflows (e.g. steps for preparing an interview or a focus group, how to run different statistical models on a software program)

4.2.3 Sampling and limitations

This study uses a convenience sampling method for data collection, recruiting researchers for whom it is convenient to participate in this study. The recruitment procedure further ensures that participants represent different domains in social sciences.

Targeted participants include senior doctoral students (in their third year or above), post-doctoral researchers, and faculty members from the Departments of Anthropology and Political Sciences and the Library and Information Science (LIS) Program at the University of Pittsburgh. A recruitment message was posted on two major social media platforms: Craigslist and Facebook. The PI of this project asked potential participants to pass along the recruitment information to others who may be interested in the research study.

For each survey profile, the participant was asked to work on 16 open-ended questions about their research data and data-sharing behaviors. They were also asked to complete 55 closed-ended questions based on the CCMF Toolkit. For each closed-ended question, the participants could freely add comments or suggest preferred exemplars that the instrument did not list. Although it might be effective to use a convenience sampling method at this exploration stage of the preliminary study, there are also several shortcomings of doing so: there might be a selection bias because all the participants are affiliated with the University of Pittsburgh and are early-careered researchers.

Four participants were interviewed (for open-ended questions) and mediated (for closed-ended ones) in July and August 2014. Each interview and mediation session was two to three hours long, allowing for a "deep dive" into scholars' data practices and capability levels. Each participant was compensated with $20-25 gift cards (USD) for their time.

Besides the interviews and mediations, the CCMF tool was emailed to a cohort of 14 participants beginning in August 2014, and nine were completed and returned as of April 2015, under a self-assessment approach. For these participants, the announced completion time was 60 minutes. Each participant was compensated $15 for their completion of the survey.

The list of participants is presented in Table 4-3.

Table 4-3. List of preliminary study participants

#	Approach	Position	Discipline	Sub-discipline
1	Interviewed and Mediated	Post-doc	Anthropology	Cultural anthropology
2	Interviewed and Mediated	Senior PhD student	Lib and Info Sci.	Music metadata
3	Self-assessed	Senior PhD student	Lib and Info Sci.	Geospatial information systems
4	Self-assessed	Senior PhD student	Lib and Info Sci.	Information retrieval
5	Interviewed and Mediated	Senior PhD student	Anthropology	Cultural anthropology, Legal Anthropology (child adoption)
6	Interviewed and Mediated	Assistant professor	Political Science	Comparative politics
7	Self-assessed	Post-doc	Political Science	Area studies (South Asia)
8	Self-assessed	Senior PhD student	Political Science	Comparative politics, political methodology
9	Self-assessed	Senior PhD student	Anthropology	Archaeology
10	Self-assessed	Visiting Scholar (Assistant Professor)	Lib and Info Sci.	Public library management
11	Self-assessed	Post-doctoral researcher	Anthropology	Medical anthropology
12	Self-assessed	Assistant professor	Lib and Info Sci.	Public library management
13	Self-assessed	Post-doctoral researcher	Lib and Info Sci.	Information seeking behaviors

4.2.4 Social scientists' data related practices

On average, participants used 6.8 words or 2.7 phrases to describe their research data. A wide range of data types are reported in

Table 4-4, with a higher proportion of observation field notes (n=8), interview records (n=8), and survey data (n=4). P01, an anthropological researcher, stated that he had been trained to collect data using a holistic approach: he usually deals with complex qualitative data, such as field notes, surveys, interviews transcriptions (categorized as interview records), maps, and material samples such as tickets or leaf samples. P03, a PhD student whose research interest is geography information systems (GIS) and accessibility, stated that her data usually has multiple attributes:

"...plus space and time. Some attributes are quantitative and some qualitative. There are often classification codes that are needed to understand some attributes" (P03).

Table 4-4. Data types (N=13)

Types of data	Freq.	%
The field notes	8	61.5%
Interview records	7	53.8%
Survey results (questionnaire)	4	30.8%
Experimental log/records	2	15.4%
Historical documents	2	15.4%
Maps	2	15.4%
Spatial data	2	15.4%
Relationship data (e.g., triples of metadata)	2	15.4%
Government statistics	2	15.4%
Participant diary	1	7.7%
Focus group	1	7.7%
Interview transcriptions	1	7.7%
Material samples	1	7.7%
Video or screencasting	1	7.7%
Archaeological field survey (excavation survey)	1	7.7%

However, political science scholars in this study handle more quantitative data. For example, P06 and P08 stated that they use government statistics and datasets for large-N analyses.

Participants were also asked about the uniqueness of their data. Nine of the 14 participants stated that their data could be fully or partially recreated and is therefore not unique. P05, a senior PhD student who studies child adoption culture in the Federated States of Micronesia, specified that regarding partial recreation:

"[In my study] legal records can be always retrieved, but I am not sure about the interview (data)" (P05).

When the participants were asked to estimate their typical data volume for one research project, the responses ranged from less than 25 MBs (n=2), 200MBs (n=1), 1-10 GBs (n=4), to

more than 10 GB (n=5). Three out of the five participants who claimed to produce more than 10 GB of data per project (P01, P02, P05) specified that their data include video, audio, photos, and screencast videos (see Table 4-5).

Table 4-5. Typical data volumes for one project (N=13)

Volume	Freaq.	Participants
Over 10 GB	5	P01, P02, P03 (5-20 GB), P05 (20GB), P10 (about 100GB)
1-10 GB	4	P04 (less than 1GB), P06 (2GB), P07 (5GB), P08 (1GB)
<25 MB	2	P11 (less than 25 MB), P13 (less than 10 MB)
Varies (hard to estimate)	2	P09, P12

Two participants answered "it depends." For example, P11 stated:

> *"In terms of computer space, very little. In terms of documents (audio files, video files, transcripts, diaries, surveys, etc.) and researcher-produced data (journal, analytic memos, code books, observation and field notes, etc.), it can be significant, especially if analyzing and coding by hand" (P11).*

Figure 4-1 summarizes the open-ended responses collected from participants. Based on the responses, it seems reasonable to conclude that social scientists have a need to reuse others' data, especially data from institutional or discipline repositories (n=10, 71%). On the contrary, only three out of 14 participants (P05, P09, and P10) had deposited their own data in repositories. Although only half of the participants had received requests to share materials or data, all participants indicated that they are willing to share upon request.

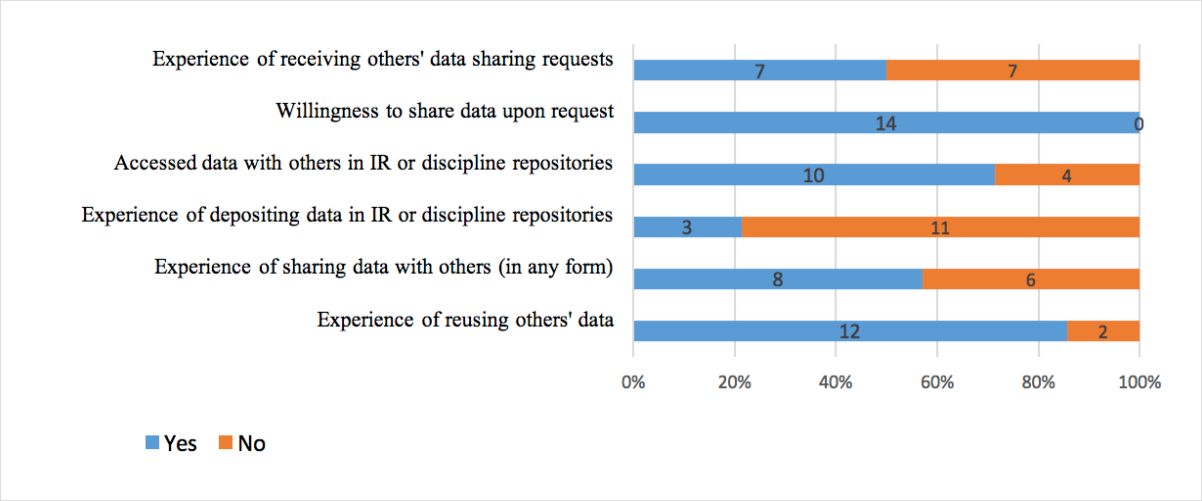


Figure 4-1. Participants' data sharing practices

4.2.5 Social scientists' data capability

Figure 4-2 presents a summary of data capability (shown in medians) in social science disciplines across all capability dimensions. The radar plot demonstrates some inter-disciplinary synergies and differences in the data-intensive capability across this section of the social sciences. For example, the dimensions of Data Common Practices and Technical Infrastructure have been highly rated by LIS researchers, even though they work in different sub-disciplines, whereas the anthropologists seem to value the dimension of Collaboration more. Political science scholars rank Legal and Ethical and Openness as the highest developments, while assigning relatively low scores to other dimensions.

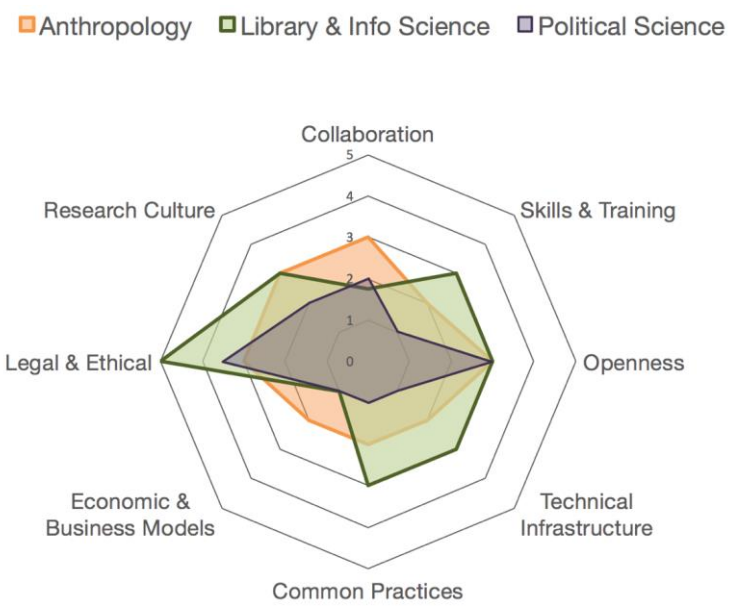


Figure 4-2. Capability summary for social sciences disciplines (by median)

As for the differences between disciplines, it is shown here that anthropology scholars' ratings were relatively evenly distributed across all dimensions. Political science scholars ranked Legal and Ethical and Openness as highest in development, whereas they assigned relatively low scores to other dimensions. LIS scholars gave better scores to Legal and Ethical but assigned higher scores to Skill and Training, Technical Infrastructure, and Common Practices than the other two disciplines.

By ranking the median and filtering the most-developed activities for each discipline, the top activities shared among two or more disciplines are identified. All items rated 3.5 or above are illustrated in a Venn diagram in

Figure 4-3, which provides a better visualization for overlapping items.

The most developed activity across the three disciplines is Openness of Published Literature. While the legal and ethical responsibilities aspect had been rated highest by both LIS and political

science researchers, in anthropology there is a mix of economic, business and collaboration concerns. Common practices related to data curation and analysis (i.e. data collection, visualization, and process workflows) are ranked higher in LIS compared to the other two fields, whereas political science's top ten list has unique items related to their openness and reuse culture.

Using the same approach, the 18 items rated in the bottom ten were visualized in a Venn diagram (

Figure 4-4). The median of all items in

Figure 4-4 are rated one (nominal activity). Data identifier, scale of infrastructure, and the use of a research discovery/networking system (e.g. CRIS) are the common items across three disciplines. The economic and business models capability dimension is most commonly perceived as weakly developed by LIS and political science.

It is worth mentioning that one of the participants also shared their "know-more moment" with us. P11 stated:

> *"The [CCMF] survey made me realize even more that we have so many technological opportunities that we aren't using and taking advantage of – especially in terms of data sharing and collaboration!" (P11)*

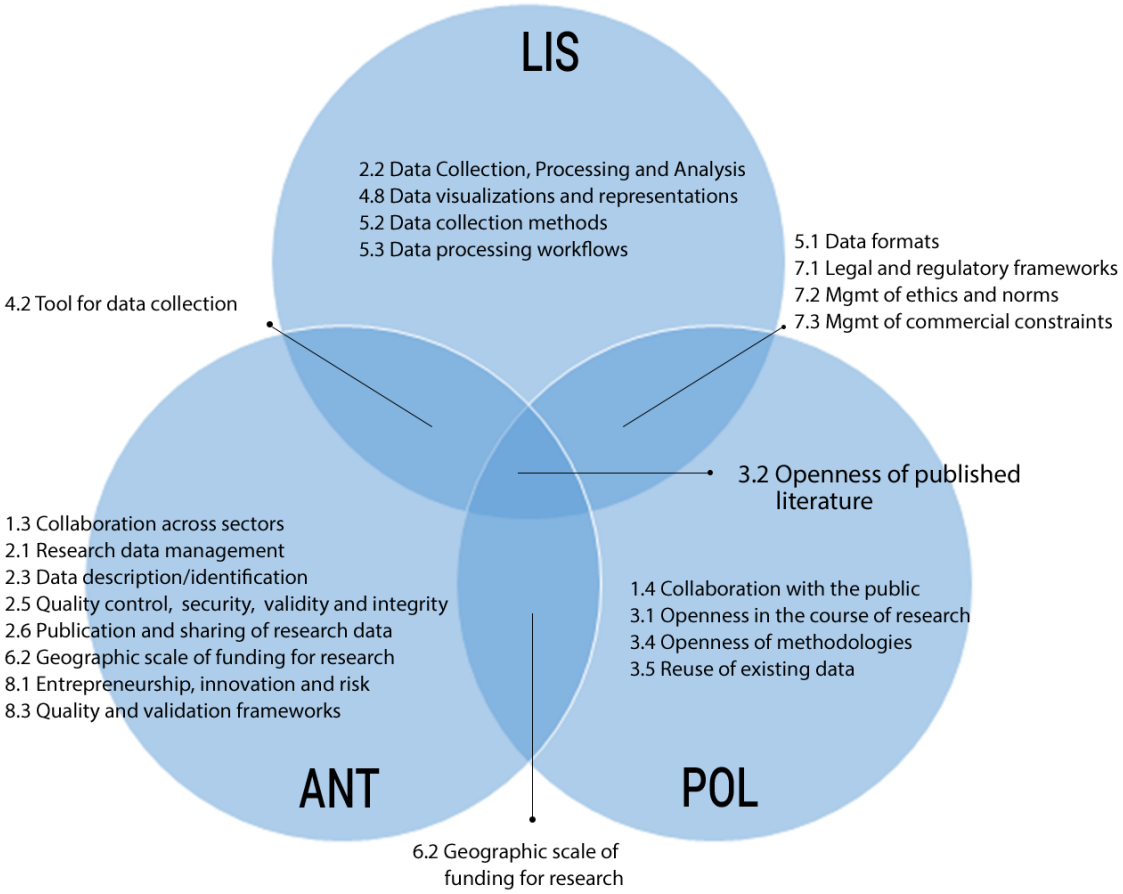


Figure 4-3. Most developed activities by discipline

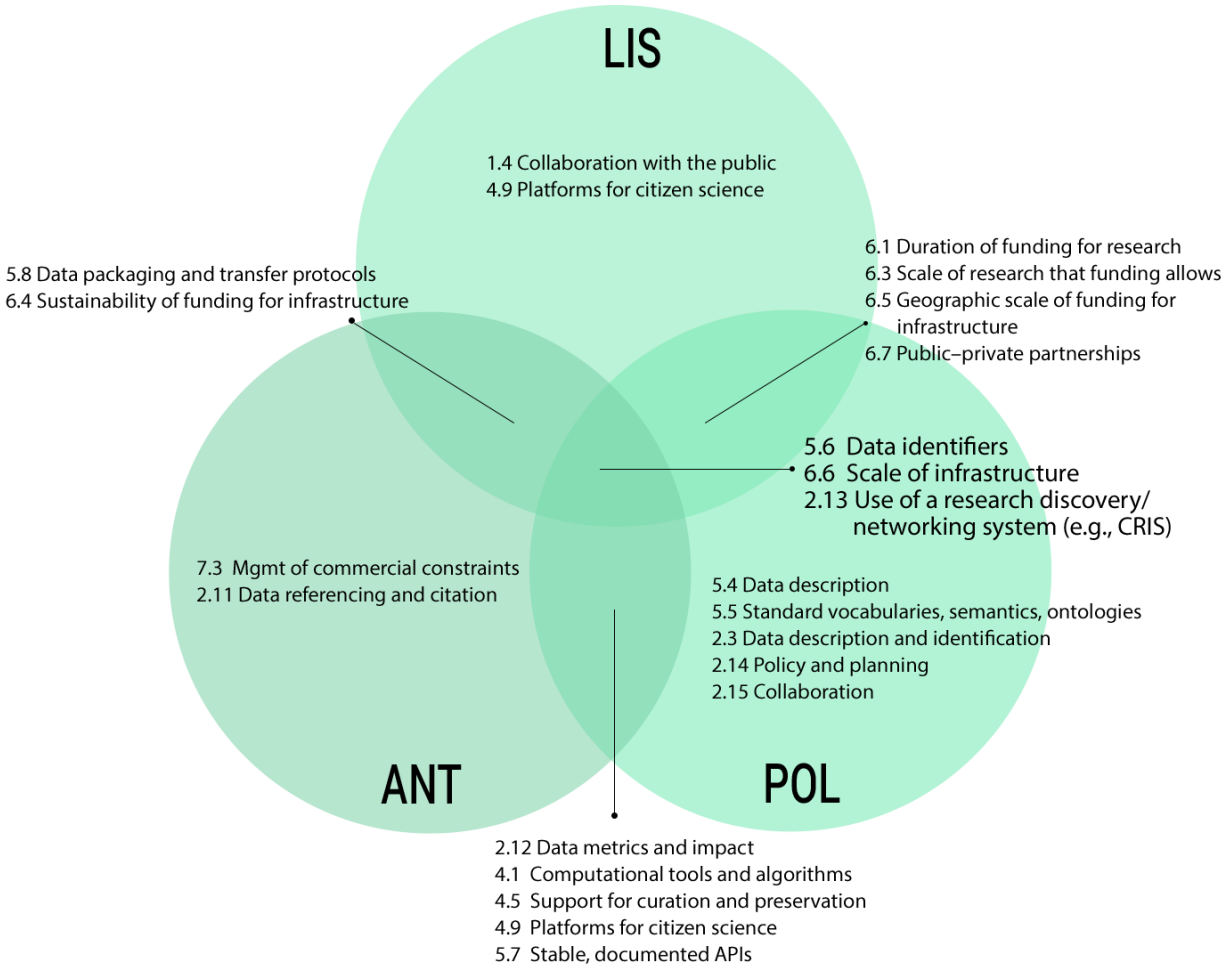


Figure 4-4. Least developed activities by discipline

4.2.6 Implications

The results from this study in select social science disciplines demonstrate the effectiveness and usefulness of the Community Capability Model tool in terms of identifying and measuring the capability for data-intensive research. However, one disadvantage of this tool is that the overall process is time consuming, making it difficult to recruit participants. The combination of open-ended questions and closed ones with commentary provides essential opportunities for the participants to extend and explain their opinions in the capability levels.

To sum up, this preliminary study confirms that the least-developed activities for social-science scholars are the 1) economic and business model, 2) skill and training activities, and 3) technical infrastructure. The preliminary results also suggest that social-science scholars have developed more maturely in legal and ethical aspects and have positive attitudes about data openness. This preliminary study informs the design of the case studies inasmuch that it is worth exploring the disciplines' similarities and deviations in data practices and capabilities.

4.3 PRELIMINARY STUDY 2: RESEARCH PROCESS STUDY

4.3.1 Research design

The visual method is an ethnographic methodology derived from visual anthropology (Pink, 2003). Today, interest in employing the visual method is growing in cultural studies, queer studies, and consumer research (Pink, 2003). Compared with purely talk-based (e.g., interviews) or text-based (e.g., diaries, social media content analysis) approaches, a visual method gathers and analyzes visual representations (Buckingham, 2009). Scholars who employ this method typically ask participants to

create video diaries, photography, or drawings as research materials. These visual products usually represent participants' life stories and points of view. The research methodology of this study draws upon visual narrative inquiry, a qualitative methodology in which participants communicate their experiences visually (Connelly & Clandinin, 1990; Bach, 2008; Bowler, Knobel, & Mattern., 2015).

This study is rooted in the observation that there are a host of visual representations of research that academic libraries use for mapping and communicating services. The study is driven by an interest to learn directly from scholars in humanities and social sciences (H&SS) about how they interact with their research data by the visualization of their own experiences.

4.3.2 Data collection and analysis

The convenience sampling method is used for the data collection with recruitment of eight H&SS scholars who have completed or were nearing completion of their doctoral degree. In December 2014 and January 2015, two focus group sessions of four participants each were held at the iSchool at the University of Pittsburgh. Seven out of eight participants held a PhD degree at the time of the study, and participants are composed of an equal number of men and women. The participants are from five research programs: information sciences, library and information science, history, anthropology, and philosophy.

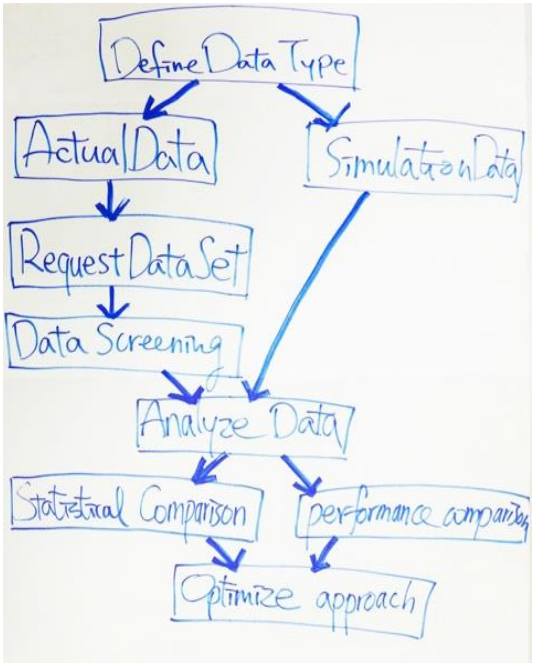
At the beginning of the focus group session, the study objectives were introduced to the participants and permission to record audio was obtained. Each participant then spent 15 minutes sketching their research process and afterwards verbally described their sketches. To minimize influencing the participants' drawings, instructions regarding how they might approach the visualizations (i.e. whether they might draw their process as a lifecycle, flowchart, or comic book panels) were intentionally avoided. The average time for each pilot focus group session is 1.75 hours.

A more detailed discussion of this research method, including the overall protocol and implications about the visual method, was reported in Mattern, Jeng, He, Lyon, & Brenner (2015).

4.3.3 Research process in humanities and social sciences

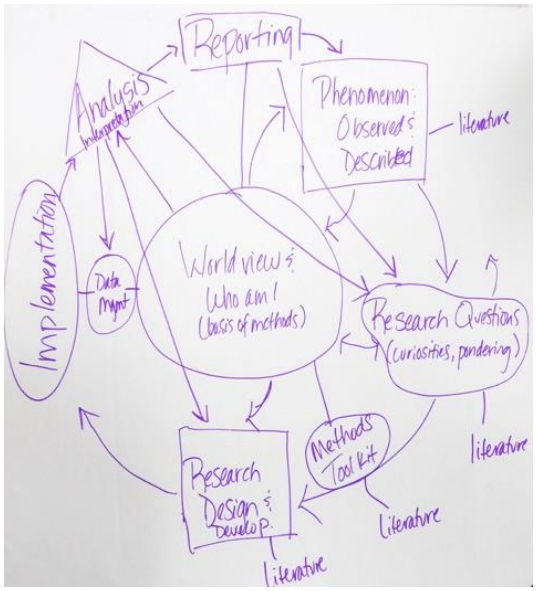
As shown in Figure 4-5, each participant in the pilot focus groups created their own sketch that visually narrates their research process and how they interact with their research data.

a. Participant No.1



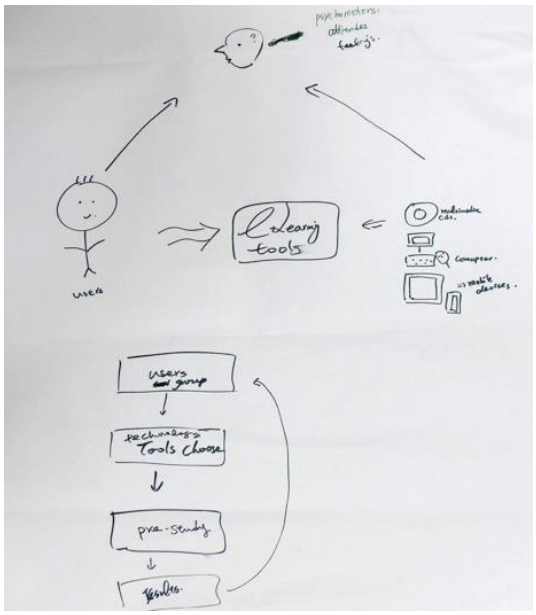
b. Participant No.2

c. Participant No.3

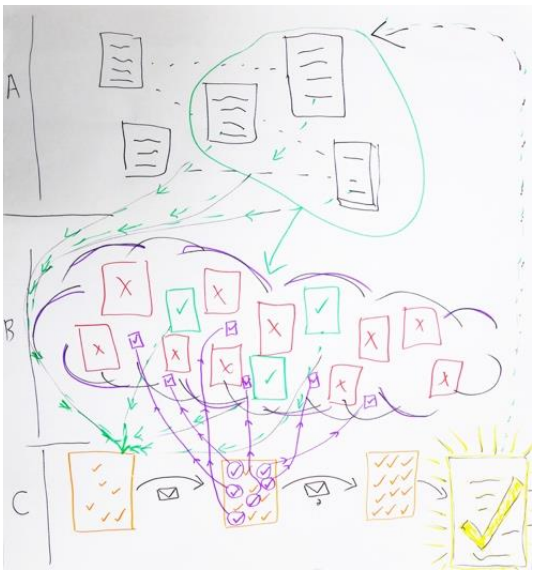


d. Participant No.4

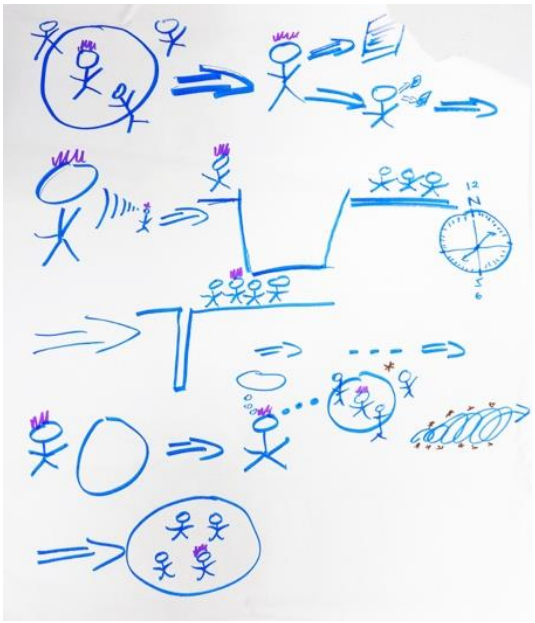
e. Participant No.5



f. Participant No.6



g. Participant No.7



h. Participant No.8

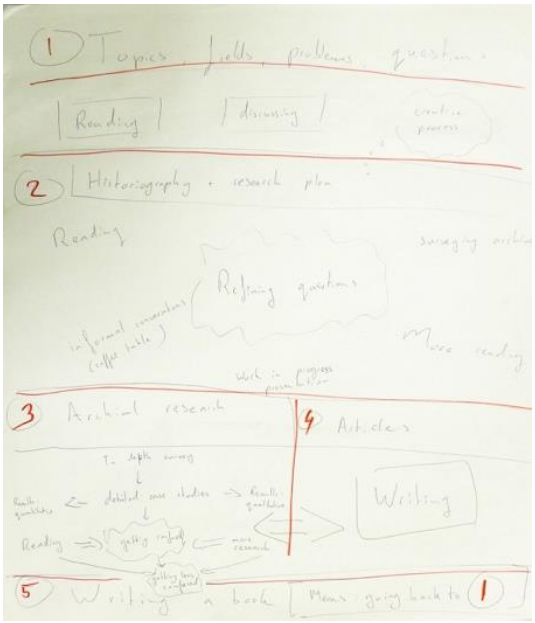


Figure 4-5. Participant 1-8 (from left to right and from top to bottom)

Figure 4-5 and Table 4-6 summarizes and compares the characteristics observed in the research processes. Most participants (i.e., P01, P02, P05, P07, & P08) depicted a linear process, displaying their work in sequential steps. P01's flowchart involves a decision about "what kind of data we're targeting, with two options: simulation versus actual data" (see Figure 4-5a). This is a variation of the linear process, one that is described in the literature (Sekaran, 2006). Other H&SS scholars visualized their research as a cyclical process. For P05 and P06 (Figure 4-5e and Figure 4-5f), when their research reaches an endpoint, they return back to the start. Their representations are reminiscent of the research processes described in Johnson and Christensen (2008).

Table 4-6. An overview of researcher participants

ID	Status	Research Focus	Institutions	Research Method/Data collection method	Observed type of visualizaiton
P01	Doctoral Candidate (ABD)	Data Mining	iSchool-IS	Quantitative/ epidemiology datasets in a historical data repository	Linear-flowchart
P02	Visiting Scholar (PhD)	Public Library Management	iSchool-LIS	Mixed/ Business documents and datasets	Linear
P03	Visiting Assistant Professor (PhD)	Public Library Management	iSchool-LIS	Qualitative/ Diaries kept by participants	Daisy
P04	Postdoctoral Researcher (PhD)	Archival Studies	History	Qualitative/Archival sources and interviews	Linear
P05	Postdoctoral Researcher (PhD)	Education Technology	iSchool-LIS	Mixed/ Survey responses, interviews	Cycle
P06	Research Fellow (PhD)	History and Philosophy of Science	Philosophy	Qualitative/ Published research articles	Cycle
P07	Postdoctoral Researcher (PhD)	Medical Anthropology	Anthropology	Qualitative/ Field notes by the researcher, interviews from informants	Linear
P08	Postdoctoral Researcher (PhD)	World History	History	Qualitative/Archival sources and Published research articles	Linear

P03 (Figure 4-5c) constructed a "daisy" shape (spoke-hub) like that described in Mackey (2009). Each of her research activities are connected to this central worldview and the resulting drawing is messier and less sequential than others produced during the sessions. For example, conducting a literature review does not occur at just one point in P03's research process. Instead, while she is preparing her research questions, selecting her methods, and designing a study, she returns to this node consistently.

While the participants' research processes might vary, this study examines them through the lens of the four main themes articulated in Table 4-7: conceptualization, planning, execution, and reporting. Note these four components are common to the research process but are not required, and there is no specific or necessarily chronological order among them. There are variations in the activities that the participants described. For example, data preparation, processing, and analysis seem to dominate participant P01's research, whereas P03 and P08 emphasize the conceptual design

of the research. Interestingly, two participants (P01 and P05) did not mention the stages of conceptualization or reporting in their sketches and narratives.

Table 4-7. Common elements of research process

				Conceptualization		Design		Execution (Interact with data)			Reporting	
#	Type	Discplines	Research Method	Developing research	Literature review	Selecting research	Defining variables,	Data gathering/ collecting	Data analysis results	Data Interpretation	Authoring	Giving access to others*
P01	Linear-flowchart	iSchool- IS	QUANT			■	■	■	■			
P02	Linear	iSchool- LIS	Mixed	■	■	■		■	■		■	■
P03	Daisy	iSchool- LIS	QUAL	■	■	■	■	■	■	■	■	
P04	Linear	History	QUAL	■	■	■	■	■	■	■	■	
P05	Cycle	iSchool- LIS	Mixed				■	■	■			
P06	Cycle	Philosophy	QUAL		■			■	■		■	■
P07	Linear	Anthropology	QUAL	■	■		■	■	■		■	
P08	Linear	History	QUAL	■	■	■		■	■	■	■	■

Note: *including publishing, sharing, and archiving to repositories.

Several interesting points were observed from the focus group outcomes. First, for P06 and P08, the literature review step includes data collection and data analysis because their main source of research data is published texts. Second, almost every H&SS scholar addressed the data gathering and data analysis steps (i.e. the execution stage of their research.) Third, only three out of eight participants mentioned giving access to others, while the rest, at most, considered the authoring step in the reporting stage. Despite the variations in research activities, the execution stage is present in every participant's sketch.

4.3.4 Research data in humanities and social sciences

Even if participants' research patterns are similar (e.g. both P06 and P07 illustrated a linear research pattern), the emphases of their research vary significantly due to the differences in their research data. This section highlights P01, P06, and P07's sketches and explores their data practices.

When primary data are from literature. P06 begins his research journey by browsing literature, represented as "Stage A" in Figure 4-5f. This is unique to the sketches, as most participants (e.g. P02, P03, P04, and P08) mentioned that they start with constructing a research idea. The cloud in "Stage B," a metaphor for a collection of useful and relevant published articles, became P06's primary material for his study. "Stage C" illustrates a submission process to an academic journal, including the interactions with potential anonymous reviewers that require a return to his data. P06 explained, "*I have to go back and find all those things that were in the cloud (Stage B)… And then I incorporate all those and send it back*." In P06's view, the articles, or "this big cloud of stuff," are the primary materials for his study. Once his article is published, his work should return and become other humanists' research materials. It is worth noting that while P06 did not explicitly mention the concept of data reuse, his sketch and narrative implicitly suggest that his published works act as data that other scholars in his community can draw upon.

When primary data are derived from informants. P07 is an anthropologist who visualized his dissertation research process in Figure 4-5g. Interestingly, he referred his trained research methodology as "a black box": "it confuses a lot of students who are in the process of doing their first fieldwork." In his sketch and narrative, he describes how he (with purple hair) gets closer to the informants as time passes: "there's a big gap, metaphorically and everything… So then after some time has passed, this gap isn't as wide and you're able to kind of leap over it and have a closer advantage with the people you're studying and interacting with" (P07). Understanding and

successfully interpreting his informants becomes the end task of his drawing. The structure of P07's research process reveals the dependency between components: a linear process implies that a step cannot begin until the previous step has finished. A "daisy"-shaped process, such as P03's visualization (Figure 4-5c), shows that the scholar might go back to the same step iteratively or might work on multiple steps concurrently.

When primary data are from a third-party repository. Although P01's sketch also shows a linear structure, her research focuses on numerical data obtained from third-party repositories, and therefore involves different activities from the others who also depicted a linear pattern. P01's primary data source is tabular epidemiology datasets from a historical data center. As shown in Figure 4-5a, the participant starts her research journey by defining the data type and splitting datasets into so-called actual data and simulation data. P01's research process is heavily driven by data and resembles a common research paradigm in recent data-intensive studies: researchers often start with data gathering, processing, and cleansing, and then try to find patterns or relationships in the data. The discovered patterns are then used to help decide where to take the research.

4.3.5 Implications

PS2's results indicate how humanities and social-science scholars visualize their research processes and data practices throughout these processes. Eight participants from five disciplines revealed different research patterns: linear, cyclical, and a daisy; as well as different focuses in their research data: from literature, from informants, and from a third-party repository.

This preliminary study contributes to the dissertation study by providing insight on diverse research patterns in the humanities and social sciences, previously relatively overlooked in the literature. Future work is needed to identify possible factors (e.g. disciplinary differences and culture, chosen research method, and primary data source) that shape individual research activities associated with data practices.

5.0 RESEARCH FRAMEWORK AND DESIGN

This chapter describes the high-level research framework, including the rationales behind selecting research methods and sampling research participants. Another goal of this chapter is to explain the relationship between the research questions and case studies. The detailed data collection and analysis of Case Studies 1, 2, and 3 are presented in Chapters 6, 7, and 8, respectively. The discussion of research findings in the three case studies is delivered in Chapter 9 (Table 5-1).

Table 5-1. A map of overall methodology and case studies

Study	Instruments	Instrument design	Data collection, analysis & results	Discussion
Case Study 1	A profiling tool	Section 5.2	Ch 6	Ch 9
Case Study 2	A survey questionnaire	Section 5.3	Ch 7	
Case Study 3	A focus group protocol	Section 5.4	Ch 8	

In the rest of this chapter, Section 5.1 provides a roadmap of the overall research design. Section 5.1.1 confirms the worldview of this dissertation, that is, the design and execution of the studies using a mixed-method approach. Section 5.1.2 explains the overall research framework, highlighting the logical connection between each case study.

Sections 5.2 to 5.4 cover the details of instrument design or development in this dissertation study, each of which is adopted in one case study:

- Instrument 1: a preliminary profiling tool for Case Study 1, designed based on the research framework;
- Instrument 2: a refined survey (a refinement of Instrument 1);
- A focus group protocol (designed based on the research framework).

A brief data collection and analysis plan is provided at the end of this chapter (Section 5.5), and the details of data collection and analysis can be found at the beginning of each corresponding chapter.

5.1 RESEARCH FRAMEWORK

5.1.1 Worldview

Creswell (2007) used the term *worldview* to describe the philosophical assumption that leads and provides the foundation for research. In the late 2000s, social scientists (e.g., Teddlie & Tashakkori, 2009; Creswell & Clark, 2007) gradually formed a consensus that *pragmatism* can provide a philosophical foundation for mixed-method research, though some scholars (e.g., Mertens, 2003) also advocate for the adoption of participatory-emancipatory research, as cited in Sung and Pan (2010).

Compared with constructivism and post-positivism that often lead to qualitative and quantitative approaches respectively, research inquiry led by pragmatism combines and synthesizes worldview components as shown in Table 5-2. Pragmatism focuses on the essence of research questions; it believes that researchers should investigate possible solutions and how to arrive at such solutions based on the nature of the problem; finally, pragmatism considers the potential effect brought by the solution (Morgan, 2014).

Table 5-2. Worldview elements

Worldview	Positivism (post-positivism)	Constructivism	Pragmatism
Ontology	External, objective, singular reality	Social and contextual, multiple realities	Reaching an external, singular reality, but acknowledging multiple subjectivities
Epistemology	Keep distance between the knower and the known	Interaction and closeness between the knower and the known	The relationship between the knower and the known depends on research phases or research questions
Nature of knowledge	Establishes verified (nonfalsified) hypotheses	Develops structual or historical insights	Accepts the variety of interests and forms of knowledge
Methodology/ Method	Deductive	Inductive	Combining, abductive (based on evidence)
	Quantitative	Qualitative	Mixed method
Example of approaches in social research	Experiments (quasi-experiments), questionnaire/scale	Ethnographic approaches, focus groups, interviews	Combining multiple approaches in a single study or multiple-study mixed method design

Source: Content in the table was collected from Creswell, J. (2013) and Kuo (2011).

The worldview in this dissertation study adheres to the philosophical foundation of pragmatism. Particularly, pragmatism helps steer the research design of this dissertation study and confirms the following three decisions:

- The methodology in this dissertation study is mixed-method, composed of three case studies. Every research approach is decided based on the research questions.
- The initial research question (i.e., RQ2B) was formed with no hypotheses, but several hypotheses were formed and described in Chapter 7.2, in terms of the factors influencing qualitative data sharing, based on data in Case Study 2.
- Data triangulation (described in Chapter 9) is constantly performed to cross validate the results. The validity of a research finding increases if it is observed repeatedly in multiple case studies.

5.1.2 Overall research design

This section describes the overall research design of the case studies in this dissertation. To address the research questions, three case studies were conducted between January 2016 and August 2016. Figure 5-1 illustrates the relationships between these case studies, and Table 5-3 serves as an index for the readers to provide a crosswalk for the research question coverage and individual case studies.

As shown in Figure 5-1, after finishing the literature review, the preliminary conceptual framework is formed. Note that two preliminary studies also contribute to the design process of the preliminary framework. The preliminary framework, formed by four dimensions (data, technology, discipline community, and individuals), is inspired by Knowledge Infrastructure and the Theory of Remote Scientific Collaboration, introduced in Chapter 3. The preliminary framework then guides the design of the instrument in Case Study 1. Specifically, it is used as a theoretical lens to provide an orienting perspective: this dissertation study explores and fills in attributes in each dimension and designs the preliminary instruments in Case Study 1.

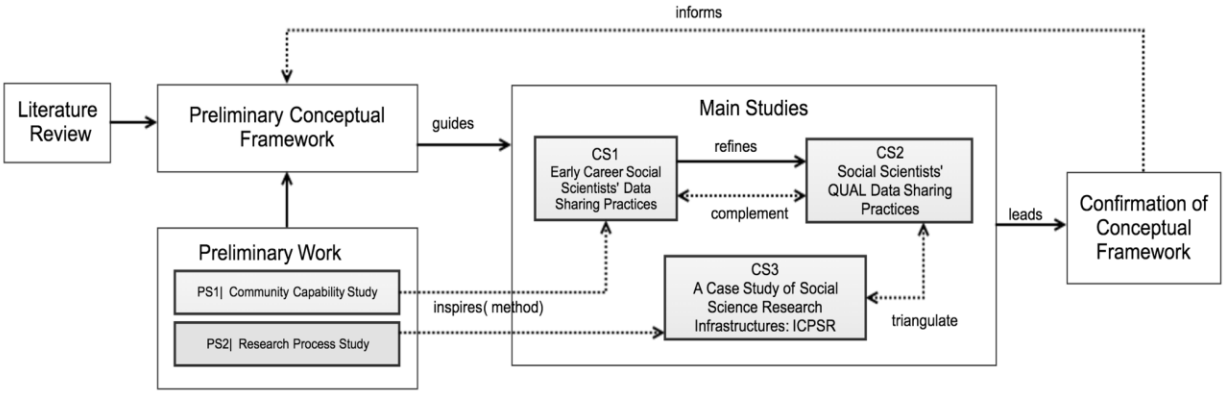


Figure 5-1. Overall research framework

Case Study 1 (hereafter: CS1) addresses the inquiry of Research Question 1 (see Table 5-3). Based on the preliminary conceptual framework, a profile tool was developed as the preliminary instrument. Case Study 2 (hereafter: CS2) then extends CS1 to address RQ2. To achieve this, the profile of CS1 was revised into an online questionnaire, which is suitable for online dissemination. It is worth noting that CS1 and CS2 can complement each other because the participants in CS1 (66 PhD early-career social scientists) have recent experience with data production but less experience with data sharing, whereas the participants in CS2 (70 senior social scientists) have experience with qualitative data sharing.

Table 5-3. A crosswalk of research questions and case studies

#	Study Alias	RQ1: Social scientists' general data-sharing practices				RQ2: Social scientists' qualitative data-sharing practices		
		1A: Research activities and research data	1B: Sharing practices	1C: Community practices	1D: Underneath technologies	2A: Sharable data	2B: Factors influencing data sharing	2C: Challenges on qual data sharing
PS1	Community cabability	◐	◐	●	●	○	○	○
PS2	Research process	●	○	○	◐	○	○	○
CS1	Early Career Social scientists DSP	●	●	◐	◐	○	○	○
CS2	Social scientists with exp. DS	○	●	◐	◐	●	●	◐
CS3	Research data infrastructure	○	○	○	●	○	◐	●

Note: DSP: data sharing practices; ● full coverage, ◐ partial coverage, ○ little or none coverage, suggesting a case study is projected to cover this sub-research question. For example, CS1 fully or partially covers the RQ1A, 1B, 1C, and 1D.

Continuing in Figure 2-1, Case Study 3 (CS3) was conducted in parallel to investigate the underlying technologies and data curation practices in a social science data infrastructure. Selecting the world's largest social science data infrastructure, ICPSR, as the study case, CS3 especially focuses on technological challenges and considerations when handling qualitative data. ICPSR employees were interviewed to obtain practical perspectives. The major role of CS3 is not only to answer the RQs, but to triangulate CS2. Since both CS2 and CS3 obtain samples via ICPSR but from different perspectives (CS2 is from the researchers' perspective and CS3 is from the perspective of data curation professionals), CS3 cross-validates CS2's result findings.

5.2 PRELIMINARY INSTRUMENT CONSTRUCTION

Traditionally, professional communities in data curation and data management fields rely on profiling tools to gather descriptions about researchers and their research data in a "concise but structured document" (Witt et al., 2009, p.3). The researchers or practitioners who use such a profiling tool can later illustrate a landscape or current state based on collected responses.

This profiling approach is shown to be useful in studying data-sharing practices, as it assists a range of stakeholders (e.g., institutions, discipline communities, and data infrastructures such as repositories or data centers) to better understand individual researchers' preparedness to share data and their actual data-sharing behaviors.

Guided by the preliminary conceptual framework, the first instrument (hereafter: Instrument 1) in this dissertation is designed as a profile tool, comprising four *dimensions* at the highest level, and then *attributes* (actual measurements; *questions* and *items* i.e., options) to examine data characteristics, technological infrastructure, perceived discipline communities, and individual characteristics and motivations to survey social scientists' actual data-sharing behaviors.

Figure 5-2 illustrates the hierarchical structure of Instrument 1: each dimension consists of several attributes; each attribute contains several questions, under which there are items. Each item is handled as an individual variable in the dataset.

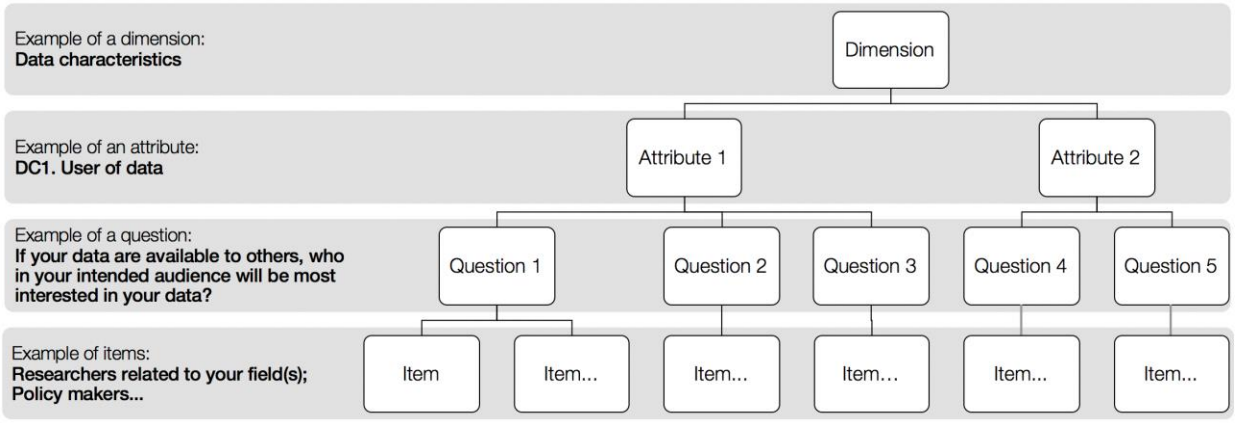


Figure 5-2. Hierarchical element of Instrument 1

Besides the four dimensions adopted from KI and TORSC introduced in Sections 5.2.1 to 5.2.4, Section 5.2.5 is appended to describe a group of questions related to social scientists' actual data-sharing behaviors.

5.2.1 Data characteristics

The nature of research data can influence the intention or decision to share. Seven attributes were developed for this dimension (Table 5-4).

Table 5-4. Dimension of data characteristics

Attributes in Data Characteristics	Examples of actual measures	Conceptual foundation or related work
DC1- User of data	Target audience of data	Data Curation Profile
DC2- Data source	Observational data, survey data, experimental data, simulation data (generated from test models)	Data Curation Profile; CCMF, University of Virginia Libraries
DC3- Data types	Text, relationship, images, or audio	
DC4- Data volume	File size, number of files in a study	CCMF
DC5- Data sensitivity	Data are sensitive or confidential	
DC6- Data shareability	Data are sharable, embargo, ambiguity of data ownership	
DC7- Data ownership	Ambiguity of data ownership	Parry & Mauthner (2004); Broom, Cheshire, & Emmison (2009)

Table 5-5. An example of customized items: data types

Items in DC3- Data types	Modified items in this dissertation	Source*
Observational data	Observational data captured in real time (e.g., fieldnotes, social experiments)	Observational
Data from informants	Data directly obtained from the study groups/informants (e.g., survey responses, diaries, interviews, oral histories)	N/A, created by this dissertation
Experimental data	Experimental data (e.g., log data)	Experimental
Simulation data	Simulation data generated from test models, where models are more important than output data (e.g., economic models)	Simulation
		Derived or compiled
Documentation-based data	Documentation-based data: records, literature, archives, or other documents (e.g., court records, prison records, letters, published articles, historical archives)	N/A, created by this dissertation
Secondary data	Secondary data (e.g., government statistics, data from IGOs or NGOs, other's data)	
Physical materials	Physical materials (e.g., artifacts, samples)	

Note: *: Data types in U of Virginia Library RDS

First, capturing *DC1- User of data* is necessary to understand social scientists' drive to share. This attribute can clarify social scientists' expectations of people who might utilize their data. For questions regarding social scientists' *DC2- Data source* and *DC3- Data types*, Instrument 1 adopts the

University of Virginia Library Research Data Services' version (n.d.) but carefully tailors it to fit the context of social science research activities. For example, in Table 5-5, four new categories are added for data type to enhance the measurement: data directly obtained from the participants, documentation-based data, secondary data, and physical materials.

Instrument 1 also captures *DC4- Data volume*. Social-science data are inherently complex and can be "big" (Dey, 1993). The volume and complexity of data (especially those involving a variety of sources) might discourage scholars from sharing (Jahnke, Asher, & Keralis, 2012). Some data might contain sensitive or copyrighted information with disclosure risks, and cannot be shared without proper handling (*DC5- Data sensitivity*).

Some data that might contain sensitive or copyrighted information also cannot be shared without proper sanitization (*DC6- Data shareability*). For example, cultural anthropologists might access sensitive marital and child adoption statuses. Another attribute, *DC7- Data ownership*, is needed to capture the ambiguity of data ownership (Parry & Mauthner, 2004).

5.2.2 Technological infrastructure

From a technical standpoint, three limitations impede the intention to share data in the social sciences (Mennes, Biswal, Castellanos, & Milham, 2013; Fecher, Friesike, & Hebing, 2015; Lyon, 2012): TI1- *Platform availability*, TI2- *Platform usability*, TI3- *Facilities*, and TI4- *Technical standards*. Table 5-6 lists these attributes and examples of their measures.

Table 5-6. Dimension of technological infrastructure

Attributes in technical infrastructure	Examples of actual measures	Conceptual foundation or related work
TI1- Platform availability	Existing disciplinary data repositories	Fecher et al., 2015; Mennes et al., 2013
TI2- Platform usability	Easy-to-use platform, tools and application's usability	
TI3- Access of tools	Access to qualitative data analysis software	Corti et al., 2014
TI4- Technical standards	Metadata standard, control vocabulary	CCMF

TI1- Platform availability examines whether there is a common, easy-to-locate platform on which social scientists can deposit data. However, even if such a platform exists, its service might not always be easy to adopt and use (Fecher et al., 2015). Related work emphasizes the importance of an easy-to-use data-sharing platform. Such a platform should contain several well-designed features, such as a simple upload mechanism or automatic data verification (Poline et al., 2012; Mennes et al., 2013).

TI2- Platform usability enables us to examine whether existing platforms are difficult to access or use due to inadequate support, e.g., the lack of access to a data analysis tool or lack of research data management resources.

Even if such platforms exist and are useful, social scientists might find them difficult to access or use due to inadequate equipment or software support (*TI3- Access of tools*), e.g., lack of skill or lack of access to the full version of a Computer Assisted Qualitative Data Analysis Software (CAQDAS), such as NVivo or ATLAS.ti. Consequently, social scientists may encounter resistance or fail to obtain support within their associated institutions. Due to insufficient technological support or associated resources, some institutes lack technical training programs or administrative support for researchers.

Finally, for each dataset shared via non-standard formats or procedures, researchers interested in reuse must investigate additional resources for interpretation. Researchers can benefit from well-defined standards that specify suggested or mandatory file formats, discipline-dependent metadata for datasets, sufficient minimal data descriptions, etc. A lack of standards could be a factor that discourages sharing and reuse. These measures are included in *TI4- Technical standards*.

5.2.3 Organization context

Table 5-7 lists the attributes related to organizational context, specifying organizational support (OC1-OC4) and a discipline community's research culture (i.e., RC1-RC4), which can influence social scientists' data-sharing practices. Based on the literature about research norms in social sciences, it seems reasonable to argue that the community plays an important role, influencing an individual's decision-making and motivation regarding data sharing.

Table 5-7. Dimension of organization context

Attributes in organizational context	Examples of actual measures	Conceptual foundation or related work
OC1- Funding sufficiency	Funding for supporting of data sharing	CCMF
OC2- Research data service (RDS) supports	Existing library RDS support	Proposed based on PS2
OC3- Internal training courses	Existing training courses	CCMF
OC4- Legal and policy	Institutional mandates	CCMF
RC1- Discipline community culture of data sharing	The culture of open sharing	Proposed by this dissertation
RC2- Discipline norms	Privacy protection for participants, continuous informed consent	Israel & Hey (2006); Israel (2015)
RC3- Research skills	Valued research skills	Proposed based on PS1 and PS2
RC4- Research activities	Research activities involved in data production	Mattern et al., 2015

Researchers may also encounter resistance or fail to obtain support within their associated institutions. Due to insufficient funding (*OC1- Funding sufficiency*) or human resources, some institutes lack technical training programs (*OC3- Internal training courses*) or research-data-related support (*OC2- Research data service supports*). Certain internal research cultures, such as unfamiliarity with appropriate methods of secondary analysis and the lack of a collaborative culture (Lyon et al., 2014), are also incompatible with sharing. Institutional policies (*OC4- Legal and policy*) about data production, management, or curation can also critically influence scholars' behavior.

To examine discipline community practices, another goal of Instrument 1 is to gather information about a community's influence on individual researchers. *RC1- Discipline community culture of data sharing* asks social scientists about their perception of community data-sharing practices. From a perspective of research norms, researchers have expressed several concerns about sharing qualitative data (*RC2- Discipline norms*). For example, some are hesitant to share qualitative data due to ethical considerations, such as continuous informed consent (Williams, Dicks, Coffey, & Mason, 2007) and the level of required privacy protection (Yoon et al., 2014; CLIR, 2013; Jahnke et al., 2012). Researchers are unsure whether they have the right to publish data or to what extent the data should be sanitized to protect participants' privacy.

Finally, inspired by Preliminary Study 2, *RC3- Research skills* and *RC4- Research activities* are included in the organizational context dimension to capture social scientists' research activities and valued skills during data production (e.g., data collection and analysis).

5.2.4 Individual characteristics and motivations

Individual characteristics such as academic position and other characteristics always play a critical role in scholars' data-sharing decisions (*IC1- Researchers' demographics*). *IC2- Cost effectiveness* is another

attribute of consideration for selective factors that influence researchers' data-sharing behaviors. Given low expected benefit or high expected effort, researchers lack incentives to share or reuse data (Kim, 2013; Kim & Stanton, 2016). Prior work identifies the challenges researchers face to provide "rich-enough" documentation of context or insufficient time to use unfamiliar data (Corti et al., 2014); Tenopir et al. (2011) also indicate that "[t]he leading reason (of why their data are not available electronically) is insufficient time" (p. 9). The attributes in the dimension of individual characteristics are summarized in Table 5-8.

Table 5-8. Dimension of individual characteristics and motivations

Attributesin individual characteristics and motivations	Examples of measures	Conceptual foundation or related work
IC1- Researchers' demographics	Prior experience, positions, etc.	Followed most recent related work e.g., Tenipir et al., 2011; 2015
IC2- Perceived ease of data sharing	Sufficient time for preparing datasets, documentation, ensuring the interoperability; administrative work	Theory of Planned Behavior ; Kim & Stanton, 2016; Wallis et al, 2013; Tenipir et al., 2011; 2015
IM1- Extrinsic motivation	Expected reward toward the career, citations	Intrinsic and extrinsic motivations from self-determination theory
IM2- Intrinsic motivation (Scholarly Altruistism)	Altruistic behaviors (e.g., sense of achievement for sharing great research)	

A lack of reward models can be viewed as a barrier for data sharing. Scholars greatly rely on a reward system in which recognitions, research funds, and credits can return to those who make contributions to creating knowledge (Kim, 2013). However, the current reward model for data sharing in the social science field is still associated with publications in formal venues (e.g., journals which receive higher SSCI impact factors). Data-sharing reward models (*IM1- Extrinsic motivation*) within social-science disciplines are still not widely recognized. Based on prior studies (e.g., Kim &

Stanton, 2016), the *IM2- Scholarly altruism* is included because altruistic behaviors strongly influence social scientists' data-sharing behaviors (Kim, 2013; Kim & Stanton, 2016).

5.2.5 Data sharing practices

Now that the dimensions that influence data sharing have been introduced, this section will elaborate the attributes that can describe the data-sharing outcome. Instrument 1 adopts an already-existing measurement (Kim, 2013; Kim & Stanton, 2016) as an outcome of social scientists' data-sharing practices. Kim's measurement covers online channels that researchers can use to give others access to their research data, as well as the frequencies in which they have done so. Since manuscripts are arguably the most common research product, the instrument also gathers information about manuscript sharing to treat as a reference point. This reference point can help further justify social scientists' research product sharing behavior.

Table 5-9. Measures in data sharing behaviors

Attributes	Examples of Measures	Conceptual foundation or related work
DS1- Data sharing (channels and frequencies)	▪ Publishing with journal venues ▪ Institutional repositories ▪ Publicly accessible web sites ▪ Academic social media platforms ▪ Discipline repositories ▪ Sent to others upon request	Kim & Stanton, 2016; Tenipir et al., 2011; 2015
DS2- Manuscript sharing (channels and frequencies)	▪ Institutional repositories ▪ Publicly accessible web sites ▪ Academic social media platforms ▪ Discipline repositories ▪ Sent to others upon request	Questions were based on DS1

Note that *TI4- Technical standards* and *TI2- Usability* were removed before carrying out Case Study 1 because it might be premature to gather detailed information about how participants assess metadata standards and the usability of data repositories, without first confirming participants' data-sharing practices.

The final version of Instrument 1 includes 99 items (appended in Appendix D):

- seven items in multiple selections,
- 88 items in multiple choice format, and
- four open-ended questions for participants who feel a need to specify their answers or express opinions outside of the closed-ended questions.

Among the 88 multiple-choice questions, 54 use a 5-point Likert scale (i.e., 1= lowest degree, 5= highest degree) which allows for future factor analysis.

5.3 INSTRUMENT REFINEMENT

There are two motives for refining Instrument 1 before conducting Case Study 2: 1) shifting the focus from RQ1 to RQ2, and 2) item reduction to create a shorter questionnaire.

First, while Instrument 1, as a profiling tool, contains a broad range of questions regarding data sharing, the refined instrument narrows the focus on qualitative data and qualitative data sharing, which addresses RQ2. In addition, researchers have pointed out that response rates drop dramatically when the announced questionnaire administration time exceeds 20 minutes (Galesic & Bosnjak, 2009).

As for the second motive, since the targeted participants in CS2 include social scientists on a national scale, the new instrument used for CS2 shall be a relatively short survey with approximately 50 items, based on the estimation of 4-5 items per minute.

Given that a refinement of Instrument 1 is essentially required, the refining process has three aims: item reduction, modification (of the remaining items), and item addition to address RQ2. Fifty-eight items are removed from Instrument 1, while 14 new items are added. The overall transformation process of Instrument 1 (with 99 items) into Instrument 2 (with 55 items) is shown in Figure 5-3. The remainder of Section 5.3 details the refinement process.

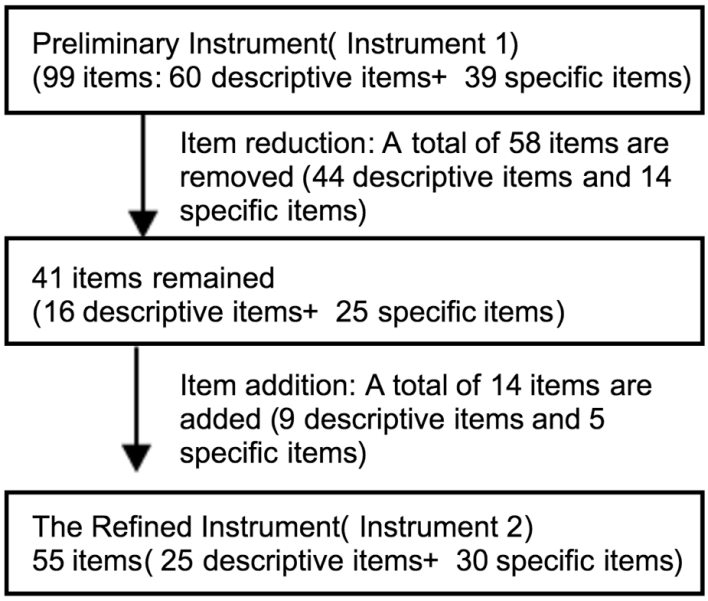


Figure 5-3. Process of instrument refinement

Note: Descriptive questions are those related to data activities and demographics; specific questions are those related to the factors influencing data sharing.

5.3.1 Item reduction

Before performing item reduction, the first step is to reorganize all Instrument 1's profile questions into two groups. The rationale behind this step is to identify items that are not suitable for a short survey or to answer RQ2. Specifically, Instrument 1 can be broken down into two categories:

- Descriptive questions related to data activities and demographics (n=60): questions designed to collect facts or tangible answers, such as demographic questions (related to researchers' attributes, e.g., age, positions, gender), data volume, primary preferred research methods, and data designated audience. Questions in this category can be flexibly adjusted according to the specific focus of RQ1 and RQ2. Note that most questions in the dimension of "data characteristics" fall into this category.
- Specific questions related to the factors influencing data sharing (n=39): potential questions that can be grouped into factors after assessments. Most questions in "individual motivations," "community culture," and "technology supports" fall into this category.

In the category of descriptive questions, 44 out of 60 items were removed or replaced, as shown in Table 5-10. For example, CS2 excludes questions about participants' data-production activities (RC4- Research activities). Also, some profiling questions such as reporting data size (DC3-Data volume) are unsuitable for the next stage.

Table 5-10. Summary of reduction of descriptive items from Instrument 1

Changed or removed items	Break-downs	# of removed items
Focus shifting	11 items in RC4- research activities 8 items in DC1- target users 5 items in DS1- manuscript sharing 9 items in RC3- research skills 7 items in OC3- of internal human supports 3 items in DC3- data volume	43
Completely replaced	1 item in DC6- shareability	1
	TOTAL	44

Note: 44 items were removed due to the change of focus from RQ1 to RQ2. One item in DC6- shareability is expended to the shareability of another seven items regarding seven types of qualitative data (e.g., researchers' notes, interview protocols…). The original DC6 question is thus removed.

For questions related to the factors influencing data sharing, an exploratory factor analysis (EFA) is used to assist the decision to reduce those items for better manageability. By using exploratory principal components analysis (PCA), and with Varimax/Orthogonal rotation and an eigenvalue cut-off of 1.0., 14 items in total are removed due to low performance in factor loading. A six-factor model was returned and explained 84% of the variance:

- perceived ease of data sharing (3 items, with 17% of explained variance),
- perceived discipline community data-sharing culture (3 items, with 16% of explained variance),
- extrinsic motivations (3 items, with 15% of explained variance),
- intrinsic motivations (2 items, with 12% of explained variance),
- perceived technical supports for data sharing-reuse (2 items, with 12% of explained variance), and
- perceived technical supports for data production (2 items, with 11% of explained variance).

Finally, six factors are obtained and listed in Table 5-11. Different researchers have reported that acceptable values of alpha should not be lower than 0.70, and higher than 0.80 is a reasonable goal (Gliem, J. & Gliem, R., 2003). Note that the Cronbach's alpha value of "perceived technical

supports for data production" is only 0.677, suggesting this dimension has relatively weak internal consistency, which could be due to a low number of questions or weak interrelatedness between items (Tavakol & Dennick, 2011).

Table 5-11. Reliability of Instrument 1 specific items

Dimension of qualitative data sharing	N	Items	Cronbach's alpha
Perceived ease of data sharing	3	▪ little effort ▪ sufficient funds ▪ sufficient time	.907
Perceived discipline community data-sharing culture	3	▪ common to see people sharing their data. ▪ there is a generic standards for data sharing. ▪ people care a great deal about data sharing.	.872
Extrinsic motivations	3	▪ help advance my career. ▪ help my publications earn more citations. ▪ give me an opportunity to collaborate with other researchers.	.817
Intrinsic motivations	2	▪ inspire other researchers or students. ▪ help others to fulfill their research need.	.822
Perceived technical supports for data sharing-reuse	2	▪ helping researchers prepare data for sharing ▪ helping researchers to reuse others' data	.856
Perceived technical supports for data production	2	▪ collecting data ▪ analyzing data	.677*

5.3.2 Item addition and Likert scale modifications

Since Case Study 2 targets participants who are likely to have experience with qualitative data sharing, 14 items were added to Instrument 2. As shown below in Table 5-12, among these 14 newly-added items, nine are descriptive qualitative-specific questions.

Five items were also added in order to 1) balance the number of items in each potential factor due to the deletion after the factor analysis, and 2) adopt new factors based on participant feedback in CS1 or recent literature. Specifically, after CS1 was conducted, recently-published literature (e.g.,

Yoon, 2016) shows that trust and confidence are associated with data-sharing and reuse incentives. Therefore, two questions are added, concentrating on the "confidence of your research" based on related work (Wicherts, Bakker, & Molenaar, 2011). Also, a "data ownership" question is added because the responses in CS1 repeatedly point out the ownership problem. Finally, "sense of good practices" is added according to user feedback in CS1.

Table 5-12. Summary of newly added descriptive items in Instrument 2

Types	Changed or removed items	Break-downs	# of newly added items
Description Questions	data shareability	▪ Detailed procedure of data collection (e.g., interview protocol) ▪ Survey instrument with actual question items ▪ Analytic scripts ▪ Multi-media ▪ Survey response (with individual responses) ▪ Interview transcripts ▪ Researcher notes	7
	data type	multimedia	1
	demographic	work sectors	1
Questions can potentially be grouped into factors	"intrinsic motivations"	provide a sample for others to learn about practicing social research methods	1
	confidence of research and data	strength of evidence confidence in the overall data quality	2
	ownership	ownership belongs to me	1
	discipline community	better sense of good practices	1
		TOTAL	14

Instrument 2 differs from Instrument 1 not only in the questions, but also in the Likert scale options. Like the design of Instrument 1, a 5-point Likert scale is used in Instrument 2 to present the better extent of the measurement. However, there are some minor modifications on Instrument 2, listed in Table 5-13. The purpose of these modifications is mainly to improve clarity of the questions, such as adding an N/A option, and to revise the midpoint option to ensure that continuous numerical

scores on a response can be obtained. The Likert scale measurement agreement (strongly disagree to strongly agree) remains the same.

Table 5-13. Modifications on Likert Scale

Types	Instrument 1	Instrument 2	Description of Modifications
Frequency	Never Rarely Sometimes Often All of the Time	Never or Rarely (about 0-10% of the time) Occasionally (about 25% of the time) Sometimes (about 50% of the time) Often (about 75% of the time) Frequently or Always (about 90-100% of the time)	Add more specific description (the description of frequency) about frequency.
Likelyhood	Very Unlikely Unlikely Undecided Likely Very Likely	Very Unlikely Somewhat Unlikely Neutral Somewhat Likely Very Likely I don't usually handle this kind of data (N/A)	Replace uncertainty midpoint response "undecided" to "N/A". The midpoint uses "Neutral" as researcher suggests (Wade, 2006)
Level of sufficiency	Not Sufficient Neutral Sufficient Not sure	Very Insufficient Somewhat Insufficient Moderate Somewhat Sufficient Very Sufficient	Change 3-point to 5-point; change neutral to moderate to ensure the response to this question as ordinal data (i.e., continuing numerical scores)
Agreement on a given statement	Strongly disagree Somewhat disagree Neither disagree or agree Somewhat agree Strongly agree	Same	--

5.3.3 Instrument 2: assessment

The final version of Instrument 2 contains 55 items, in which 25 are for descriptive statistical analysis, and 30 items will be tested by a principal component analysis (PCA) in Case Study 2.

This section presents the assessment in terms of the reliability of Instrument 2 based on CS2 responses (n=70). By using exploratory principal components analysis (PCA), and with Varimax/Orthogonal rotation and an eigenvalue cut-off of 1.0, a seven-factor model was returned

and explained 73% of the variance. The factor loading table is provided in Table 5-14. It is important to note that the factor loading suggests merging the following two dimensions together:

- perceived technical supports for data sharing-reuse, and
- perceived technical supports for data production

After merging these two technical-support dimensions, the item "reuse others' data" is revised to "discover others' data" in order to increase the readability of this item, because the term *reuse* may comprise several activities such as discovering, accessing, and re-analyzing (Curty & Qin, 2014).

Table 5-14. Factor loadings

Rotated Component Matrix[a]	Component						
Factor loadings	1	2	3	4	5	6	7
sufficient time	0.401	0.285	0.043	0.704	0.082	0.078	-0.124
little effort	0.057	0.082	0.136	0.867	0.068	-0.133	0.083
sufficient funds	0.086	-0.08	-0.014	0.784	0.029	-0.077	0.159
tech for analyzing data	0.069	0.46	-0.086	0.028	0.601	-0.162	0.105
tech for collecting data	0.216	0.072	-0.094	-0.07	0.756	-0.064	0.131
tech for discovering others' data	0.053	0.284	0.099	0.143	0.667	0.267	-0.203
tech for preparing data for sharing	0.007	-0.034	0.084	0.141	0.824	0.21	-0.187
common to see people sharing their data	-0.035	0.018	0.015	-0.126	0.151	0.843	0.015
there is a standard for data sharing	-0.047	0.089	0.064	0.03	0.065	0.727	0.223
people care a great deal about data sharing	0.222	0.144	0.132	-0.079	-0.049	0.733	0.002
inspire others	0.148	0.805	0.221	0.009	0.193	0.191	-0.055
fulfill their research need	-0.151	0.747	0.234	0.131	0.059	0.199	0.083
provide a sample for learning methods	0.021	0.838	0.185	-0.022	0.123	-0.017	-0.099
collaborate with other researchers	0.095	0.328	0.71	0.099	-0.157	0.037	0.237
help earn more citations	-0.026	0.127	0.877	0.121	0.043	0.175	0.084
help advance career	0.225	0.206	0.875	-0.022	0.066	0.018	-0.015
appropriate reused	0.728	-0.087	0.144	0.32	0.077	0.093	0.275
appropriate interpreted	0.694	-0.22	0.141	0.442	0.051	-0.039	0.121
confidence in the overall data quality	0.764	0.116	0.061	0.063	0.061	-0.068	-0.117
strength of evidence	0.854	0.069	0.015	-0.046	0.14	0.15	0.045
data belongs to me	-0.022	-0.081	-0.033	0.173	-0.116	0.155	0.843
complete rights	0.15	0.034	0.296	0.022	0.022	0.076	0.81

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.
Seven factors have been extracted. Can explain 73% of variance.

Values of Cronbach's alpha in Table 5-15 are used to assess both reliability and internal consistency of items. Cronbach's alpha enables the measurement of the degree to which different items are correlated and the measurement of internal consistency. All the variables are above 0.70, suggesting they all have acceptable (>0.70) or fairly good (>0.80) internal consistency (Gliem, J. & Gliem, R., 2003).

Table 5-15. Reliability assessment in Instrument 2

Variables	Number of items	Items (item to all correlations)	Cronbach's alpha	% of variance
Trust of data quality and being reused	4	▪ strength of evidence (.854) ▪ confidence in the overall data quality (.764) ▪ appropriate reused (.728) ▪ appropriate interpreted (.694)	.824	12.4%
Intrinsic motivations	3	▪ provide a sample for others to learn methods (.838) ▪ inspire other researchers or students (.805) ▪ help others to fulfill their research need (.747)	.832	11.7%
Extrinsic motivations	3	▪ help advance my career (.875) ▪ help my publications earn more citations (.877) ▪ give me an opportunity to collaborate with other researchers (.710)	.849	10.8%
Effortless of sharing	3	▪ little effort (.867) ▪ sufficient funds (.784) ▪ sufficient time (.704)	.767	10.5%
Tech supports	4	▪ collecting data (.756) ▪ helping researchers prepare data for sharing (.824) ▪ analyzing data (.601) ▪ discover others' data (.667)	.745	10.1%
Discipline community practice	3	▪ common to see people sharing their data (.843) ▪ there is a generic standards for data sharing (.727) ▪ people care a great deal about data sharing (.733)	.723	9.7%
Data ownership	2	▪ the ownership belongs to me (.843) ▪ complete rights (.810)	.744	8.0%

5.4 FOCUS GROUP PROTOCOL DESIGN

Though data curation and data processing are important aspects of completing the research data-sharing process, there are few third-party studies examining how data curation practices work in social sciences. Therefore, a study is needed to address this.

Case Study 3 (CS3) adopts a focus-group approach to interview curation professionals and other professionals at a research data infrastructure, ICPSR. The rationale for using a focus-group approach is to draw upon participants' experiences and encourage interaction among group participants.

CS3 was conducted in parallel with Case Studies 1 and 2 (Figure 5-4). Since most of the questionnaire questions in CS2 are closed-ended, participants might be limited when describing their qualitative data-sharing experiences and needs. Hence, CS3 was performed to allow participants to reflect on underlying technologies and challenges they face when depositing data at ICPSR, thereby strengthening the research outcomes of RQ1 and RQ2.

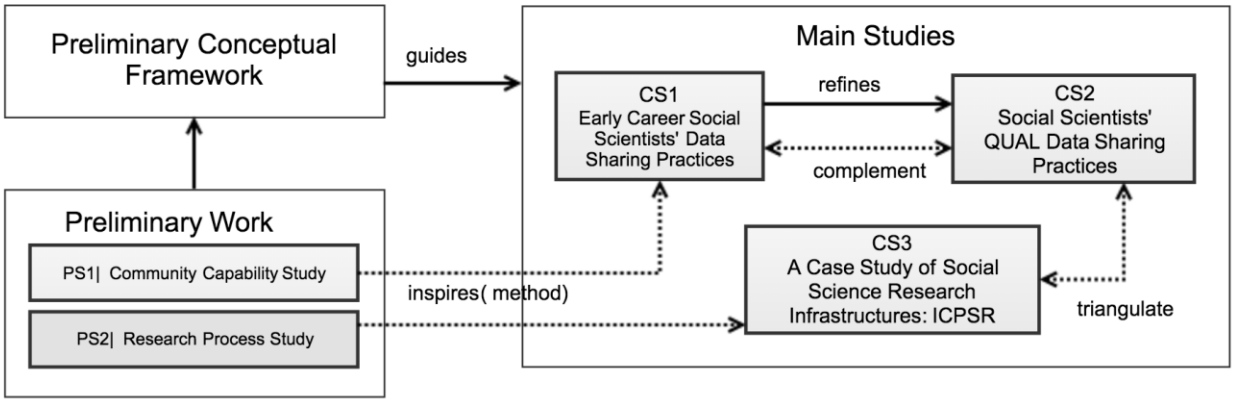


Figure 5-4. A closer look at relationships between studies (extracted from Figure 5-1)

Directly inspired by the previously-conducted Preliminary Study 2 (Figure 5-4) in this dissertation study, the detailed execution of the focus group follows Mattern et al. (2015) and Lyon et al. (2017) via a visual narrative inquiry technique. In particular, this study uses a visual approach that asks participant to write down important concepts on sticky notes, then place and sort them to create a group outcome. The sticky note technique is believed to enhance interaction to "draw out reluctant participants, and help create a group outcome" within focus group participants (Peterson & Barron, 2007, p.140). Specifically, in Lyon et al. (2017), researchers asked focus group participants to write down specific actions related to ensuring research transparency. Each participant was asked to place notes in a data lifecycle diagram. Later, sticky notes were kept as physical data recorded by researchers.

The study introductory script shown to participants and the details of the focus group protocol is attached in Appendix G and Appendix H.

The brief study protocol, as shown in Table 5-16, begins in *Stage I*: study information is introduced to and consent is obtained from the focus group participants. Participants are then invited to describe their backgrounds and explain how their backgrounds have led them to their job positions at ICPSR.

Table 5-16. Process of the focus group design

Stages	Description
I. Warming up	▪ The mediators introduce the study information and inquire consent. ▪ Participants describe their background and explain how their backgrounds led them to their current job positions.
II. Session of professional activities	▪ Each participant writes down their professional activities (related to their responsibilities at their institution) regarding data curation or collection development at the institution, one activity per sticky note. ▪ All participants leave the table and go to the whiteboard, self-grouping the sticky notes they have. Participants may use magic markers as a visual aid or re-position the sticky notes.
III. Underneath information technology activity-collecting *ITs and desired ITs*	▪ Participants are back at the table and, on another set of sticky notes, write down the tools related to the concepts on the whiteboard, such as certain software, online services, or homegrown programs. ▪ Participants describe desired information technologies.
IV. Follow-up questions	Each participant elaborates more about their actions in curation, acquisition, and collection development.

Note: The detail procedure is attached in Appendix H

In *Stage II*- Session of Professional Activities, each participant writes down their professional activities (related to their day-to-day responsibilities at their institution) regarding data curation or collection development at the institution, one activity per sticky note. The participants then have a discussion among themselves and explain these activities to each other. Next, they work on sorting these actions into clusters. They are encouraged to leave their seats and go to the whiteboard, self-grouping their sticky notes. They may use magic markers as a visual aid or re-position the sticky notes as they see fit.

In *Stage III*, participants are sent back to the table and, on another set of sticky notes, write down the tools related to the sorted concepts on the whiteboard, such as certain software, online services, or homegrown programs. Participants are then encouraged to describe imaginary or desired information technologies.

In the final stage, participants are asked to elaborate challenges and opportunities regarding data-sharing practices, as well as additional questions about ICPSR's professional activities. While Appendix H lists all questions, here are some examples of them in *Stage IV*:

- Please elaborate more about the differences between curating qualitative, mixed-method, and quantitative data, if any.
- What are critical factors that may influence researchers' willingness to share their data?
- How do you determine the scope of ICPSR's collection?
- Does ICPSR provide other services or support to further connect the data depositors and data reusers?

Data collection and results are reported in Chapter 8. Participants are not given any hints about pre-defined frameworks, nor were they limited on how activities should be organized during the focus group sessions. The rationale behind this neutral setting is to capture real practices and participants' perceptions without undue influence.

5.5 SAMPLING RATIONALES AND DATA ANALYSIS PLAN

This section reports the rationales for choosing sampling methods and the data analysis plan for each case study.

5.5.1 Sampling rationales

5.5.1.1 *Case Study 1*

Case Study 1's participants are targeted using total population sampling (
Table 5-17). Total population sampling is one approach of purposive sampling, which is based on a specific purpose compared with random sampling (Teddlie & Yu, 2007; Etikan, Musa, & Alkassim, 2016). However, targeting all PhD students and post-doctoral researchers in the country is practically impossible. Therefore, to reach an accessible population, a convenience sampling method was used by inviting all PhD students and post-docs at the University of Pittsburgh and Carnegie Mellon University, U.S. The rationale for targeting early-career researchers is that they tend to be engaged in every research stage or all activities of their own dissertation projects, including data collection, processing, and analysis, whereas senior researchers might focus more on high level decision-making such as grant writing, constructing ideas, and interpreting data. The target population includes all currently-enrolled (at the point of January 2016) PhD students and post-doctoral researchers in 20 department or academic units at the University of Pittsburgh and four department at Carnegie Mellon University (CMU).

Table 5-17. Summary of case study participants and sampling rationales

	Case Study 1	Case Study 2	Case Study 3
Target population	Social scientists who are involved in most stages of data production and sharing, e.g., PhD students and post-doctoral researchers	Social scientists who have qualitative data-sharing experience at discipline repositories	Directors or staff at discipline repositories
Sampling methods-targeting populations	Targeting method: Purposive sampling (total population sampling)	Targeting method: Purposive sampling (sampling to achieve repetitiveness)	Targeting & access method: Purposive sampling-Expert sampling
Sampling method-accessible samples	Accessible method: Convenience sampling approach - captive sample	Accessible method: Purposive sampling - Critical Case Sampling (accomplished by performing a keyword search)	
Final samples	PhD students and post-doctoral researchers at PITT and CMU	Data depositors at ICPSR and QDR, whose study description contains qualitative methods	Directors or staff in research data infrastructure at ICPSR
Researchapproach	Survey-Online questionnaire	Survey- Online questionnaire	On-site focus groups and an interview

5.5.1.2 *Case Study 2*

To achieve repetitiveness, the target population in Case Study 2 is social scientists who have experience with qualitative data sharing. Two platforms, ICPSR and QDR, are selected because ICPSR is the largest social-science repository and QDR is one of the few repositories that stores qualitative data in social sciences. One foreseen challenge is to identify individuals with qualitative research experience at ICPSR. Subsequently, this dissertation study takes advantage of the dataset keywords on ICPSR and identifies potential PIs by performing qualitative-research-relevant keyword searches within the past ten years. This ten-year timeframe ensures that the most recent status of PIs is reflected. After eliminating duplicate study entries from the keyword results, PIs' names and contact information were extracted and the questionnaire invitations were sent to all potential participants.

5.5.1.3 *Case Study 3*

The sampling method in Case Study 3 is expert sampling, targeting data-curation professionals and other professionals who work at ICPSR. To contact such a specific target population, the invitations are sent according to the ICPSR staff directory or are referred by ICPSR employees.

5.5.2 Data analysis plan

Case Study 1 focuses on a holistic view of the current state of data sharing in social sciences. Instrument 1 provides a blend of quantified descriptions on possible variables (e.g., frequency of data sharing), countable categorical results (e.g., data types), and qualitative descriptions (e.g., research interests and comments). Descriptive statistics are used in a large proportion of Case Study 1 to portray basic characteristics. Inferential statistics, such as ANOVA and nonparametric tests, are used to compare different means or distributions within a specific group. For example, an ANOVA is used to discover there is no significant difference among the three research method groups (QUAL, QUANT, and MIX) in terms of reported data size.

In Case Study 2, both descriptive and inferential statistics (e.g., nonparametric tests, multi-level analysis, etc.) are used to summarize the dataset, compare results with CS1, and determine the predictors of data-sharing behaviors. Both Case Studies 1 and 2 involve statistical software packages and spreadsheet-style tools (see Table 5-18).

Table 5-18. Tools help data production in this dissertation study

Studies	Processing tools	Analytic tools	Visualization tools
Case Study 1	SPSS, Excel	SPSS	SPSS, Excel, Tableau
Case Study 2	SPSS, Excel, Qualtrics, Homegrown Python scricts	SPSS	SPSS, Excel, Tableau
Case Study 3	Paid transcribing services	ATLAS.ti, Excel	Photoshop, Gephi, ATLAS.ti, Voyant (text mining tool)

The data collected in Case Study 3 are essentially qualitative: physical sticky notes, photos of visualizations that participants created during the focus group, and audio files recorded during the interview and focus groups. After collecting data from the research sites, all the sticky notes are digitalized and entered into a spreadsheet-style table. The audio files are transcribed to text-form data by a paid service (iScribe). Participants' quotations on transcription files are managed using ATLAS.ti, a qualitative data analysis software package.

5.5.3 Data triangulations

Data triangulation involves the processes that use "different sources of information in order to increase the validity of a study" (Guion, Diehl, & McDonald, 2011, para 3). According to Olsen (2004), triangulation in social research not only serves to increase validity, but also to deepen and widen researchers' understanding and "support interdisciplinary research" (p.1). These approaches usually start by identifying different stakeholder groups (e.g., data depositors and data curation professionals at ICPSR). After summarizing the research findings in each case study (Chapter 6, Chapter 7, and Chapter 8), this dissertation study compares them to identify agreements or divergences. The results are discussed in Chapter 9.

6.0 CASE STUDY 1: EARLY-CAREER SOCIAL SCIENTISTS' DATA-SHARING PRACTICES

6.1 OVERVIEW OF CASE STUDY 1

This case study investigates the landscape of data-sharing practices in social sciences using Instrument 1 in this dissertation study. To ensure that the preliminary instrument can be applied in real and practical terms, a case study is conducted by collecting responses from 93 early-career social scientists at the University of Pittsburgh (PITT) and Carnegie Mellon University (CMU), U.S.

The results suggest there is no significant difference among early-career social scientists who prefer quantitative, mixed, or qualitative research methods in terms of research activities and data-sharing practices. In addition, this study confirms that there is a gap between participants' attitudes about research openness and their actual sharing behaviors, highlighting the need to study the "barriers" along with the "incentives" of research data sharing.

6.2 DATA COLLECTION

Survey invitations were sent to 553 potential participants in 20 social-science-related units (Appendix A) at two universities. Among the invitation emails sent to PITT participants (498 out of 553), 17 were

immediately rejected by the email service system, possibly due to account expiration after users left the organization.

With an online questionnaire link (Qualtrics), an invitation to complete the profile was sent in December 2015, and a reminder was sent in February 2016. This collection process received responses from 93 out of the 536 successfully-delivered invitations, resulting in a 17.4% response rate. This rate is highly comparable to that of related work (with response rates of 9-16%) (Kim & Stanton, 2016; Tenopir et al., 2011). Among the 93 responses, 66 completed the full profile. These 66 completed profiles were the final samples used in this study. After removing two extreme values (23.4 hours and 8.82 hours), the average survey completion time for the remaining 64 participants is 13.4 minutes.

6.3 RESULT FINDINGS

6.3.1 Research activities

Table 6-1 summarizes the distribution of the sampled participants by preferred research method and discipline groups. Both *Policy & Political Science* and *Education* have a non-negligible portion favoring QUANT and MIX approaches. Participants in *Economics & Business* overwhelmingly select QUANT approaches as their preferred method. *Information & communication* participants identify MIX approaches as their method of choice.

Table 6-1. A cross-tabulation of preferred research methods and disciplines

		Self-identified preferred research methods			TOTAL
		QUANT	MIX	QUAL	
Discipline Groups	Eco & Business	12	1	0	13
	Info & Communication	1	5	2	8
	Policy & Political Sciences	7	6	0	13
	Psychology & Decision sciences	12	2	0	14
	Education	7	4	0	11
	Sociology & social work	1	0	4	5
	History	0	2	0	2
Total		40 (60.6%)	20 (30.3%)	6 (9.1%)	66

The research findings reveal how frequently the participants in each method group (i.e., QUAL, MIX, and QUANT) perform individual research activities. These research activities include

Planning, Literature Review, Data Gathering, Data Processing, Data Analysis, Result Interpretation, Authoring, Publishing, and Data Reuse (Mattern et al., 2015).

Figure 6-1 summarizes the results of the research activities involved in participants' general research work, where legends ★, ▲, and ○ represent qualitative, mixed, and quantitative groups, respectively. Participants are asked to what extent certain research activities might be involved in their research. Frequency is measured on a scale from 1 (never) to 5 (all the time). The light blue band indicates the range (difference) among observed values.

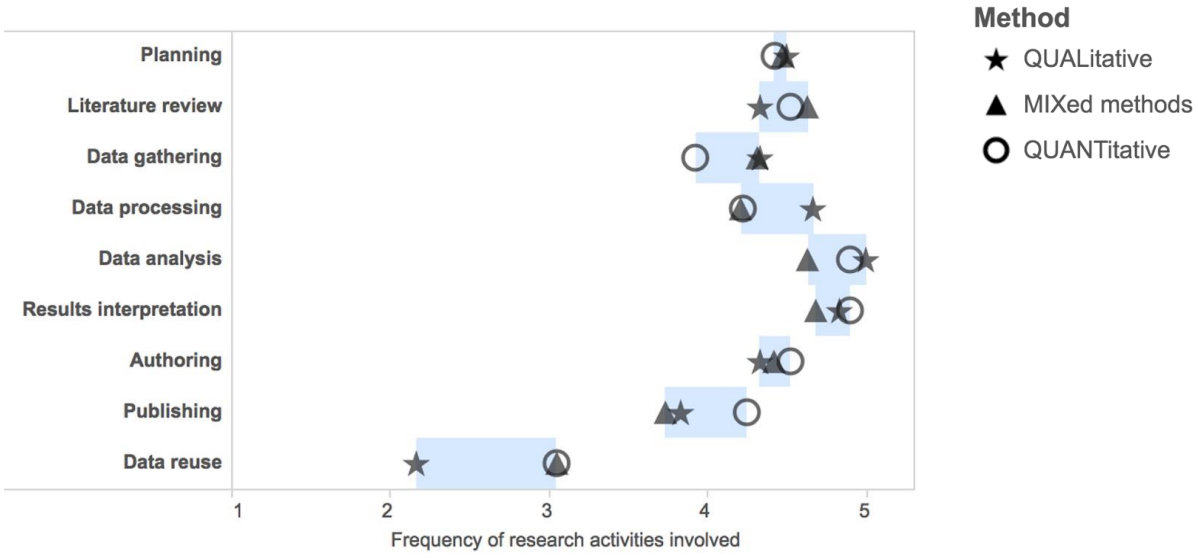


Figure 6-1. Research activities involved in social scientists' general research projects

There are several interesting findings. First, counterintuitively, there is no significant difference between qualitative and quantitative methods, even for data-related activities such as data processing and analysis. There is a significant difference between the frequencies of data analysis on different research methods at the $p<0.05$ level conditions [$(2, 62) = 4.32$, $p=0.018$]. Post hoc

comparisons using the Tukey HSD test suggest that the mixed-method approach (M = 4.63, SD = 0.114) is significantly lower than the other two.

Second, the MIX group does not always fall between QUAL and QUANT—an interesting pattern worthy of future investigation. Different averages are also observed in the "publishing" and "data reuse" stages. A subsequent ANOVA test suggests that researchers whose primary method is QUANT report more frequent publishing activities than the other two methods.

6.3.2 Research data characteristics

For social scientists' research data, this section reports results from four research data characteristics: data volume, data type, whether the data can be shared, and the intended audience of the data.

Among the 61 participants who completed the survey and reported data volume, two-thirds deal with data on the scale of megabytes (N=44), thereby confirming that they are working on small-data rather than big-data projects. Specifically, 26 participants report volumes between 0-100 MB, 18 report 100 MB-1 GB, 15 report 1 GB-10 GB, and five report to having more than 5 GB. The average data volume is 4.25 GB per research project, with a median of 200 MB, indicating the existence of outlier values much higher than the average. Although the majority (61 out of 66) report an estimated size, there are still five participants who answered "unknown." In the Implication of Instrument section (10.1.2), reflections are discussed for how future work can modify this kind of question and further improve the response rate.

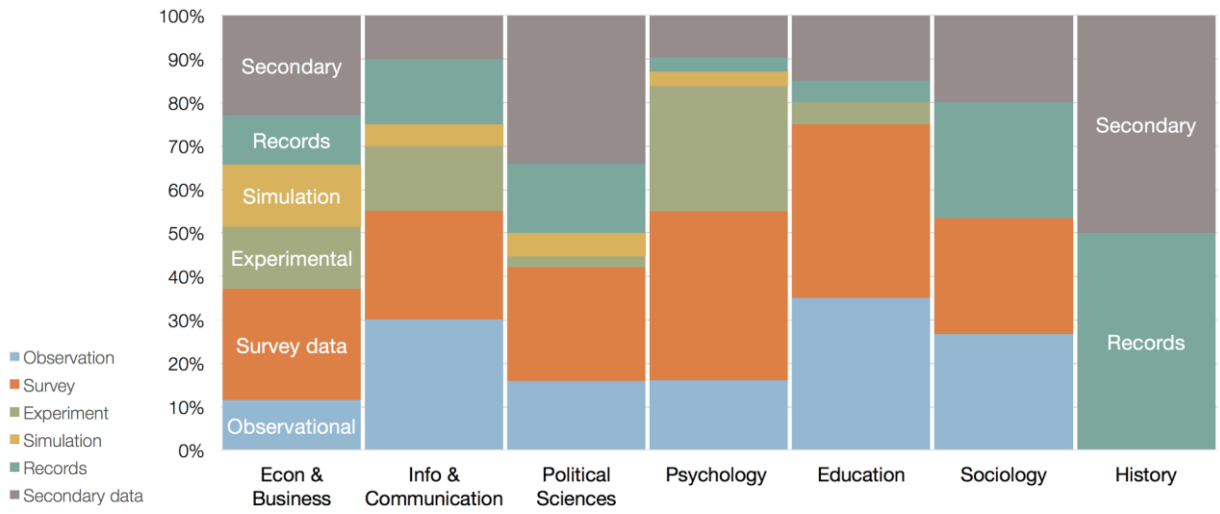


Figure 6-2. Data types and discipline categories

The average data volume of QUANT projects is 5.4 GB, much larger than that of QUAL (2.6 GB) or MIX (2 GB). However, through an ANOVA, there was not enough evidence to support the hypothesis that there is a significant difference of data volume among these three research methods.

Figure 6-2 illustrates the distribution of data types in each discipline. Although Economics is biased toward QUANT in terms of a primary research method, its data type is diversified. The data type reported by Education, Sociology, and History researchers are less diverse and centered on qualitative data, such as records and observational data.

This case study further investigates whether research methods are associated with shareability of research data. When asked if their data is sharable, the majority of participants report that their data is completely shareable (N=14, 21.2%) or mostly shareable (N=28, 42.4%). However, about 5% of participants think their data is not allowed to be shared. Table 6-2 summarizes the answers reported by participants in the different method groups. Although the QUAL group

appears to skew toward "not shareable" compared with the QUANT and MIX groups, the difference is not statistically significant in a chi-square test, where χ^2 (4, N = 61) = 8.92, p=0.06, at the 0.05 level. Note that because the chi-square test requires the expected value in each cell to be greater than 5, the analysis only includes data for Completely sharable, Mostly sharable, and Partially sharable.

Table 6-2. A cross-tabulation of data shareability and research methods

	Preferred methods			Total
	Quant (n=40)	Mix (n=20)	Qual (n=6)	
Completely Sharable	10	4	0	14
Partially Sharable	17	10	1	28
Partially Sharable	9	5	5	19
Not allowed to share	2	1	0	3
Other	2	0	0	2

As for the target audience for the data, "researchers in the same discipline" wins by a landslide, mentioned by 93.9% (62 out of 66) of the participants. In second place, surprisingly, is "graduate students" (40 out of 66, 60%), suggesting that participants perceive the value of teaching and learning from research data. The third and fourth are the practitioner (25 of 66, 37.9%) and policy maker (25.8%), respectively. Government administration, research participants, and researchers outside the field are also mentioned by over 20% of participants. Note that the participants are allowed to select more than one target audience, so the total exceeds 100%.

6.3.3 Current practices of data reuse and sharing

Figure 6-3 reports the frequency of sharing data in the past three years on five channels, including Institutional Repositories, Public Websites, Academic SNS, Discipline Repositories, and Via Emails. The frequency is scaled between 1 (never) and 5 (all the time). In an attempt to establish a meaningful baseline, the profile instrument also asked about the frequency of sharing manuscripts (including pre-prints) in addition to sharing datasets, because manuscripts can be seen as the most commonly generated research product.

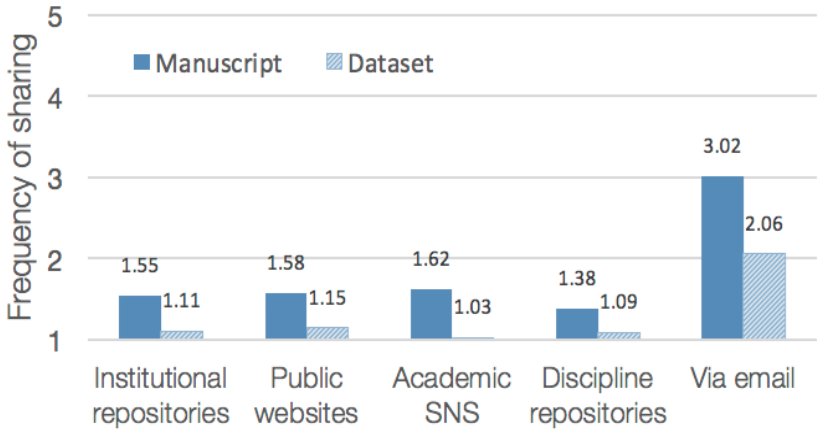


Figure 6-3. Frequency of sharing research products on five sharing channels

Unsurprisingly, the frequency of manuscript sharing is slightly higher than that of dataset sharing. However, sharing frequency remains consistently low for the five channels and the two types of research products. Before manuscript sharing becomes a common practice, it might be difficult for researchers to take the additional step toward dataset sharing. To validate this hypothesis, further investigation is needed to explore the relationship between the frequency of data

sharing and preprint sharing. An ANOVA test was executed to evaluate the disparity of data-sharing frequency among different participant groups. Similar to research activities, the averages of data-sharing frequency are consistently low across the three participant groups (i.e., researchers who preferred QUAL, QUANT, and MIX methods) without a significant difference.

6.3.4 Perceived discipline community culture

Table 6-3 shows a list of items in discipline community cultures, where 1 represents strongly disagree and 5 represents strongly agree. The majority of participants (strongly or slightly) disagree with the existence of a standard procedure and well-known, recognized data infrastructure. The result is consistent with PS1's findings that standards are one of the least-developed capabilities in social science disciplines.

Table 6-3. Perceived community culture

Community culture	M	SD	1	2	3	4	5
common to see people sharing their data.	2.92	1.154	11.30%	27.40%	17.70%	32.30%	9.70%
there is a standard for data sharing.	2.11	1.149	36.80%	33.30%	15.80%	8.80%	5.30%
people care a great deal about data sharing.	2.88	1.223	15%	26.7%	21.7%	28.3%	8.3%

Note: each question is preceded by a context description: "Please answer the following questions about your discipline community regarding research data sharing."

6.3.5 Institutional and technological supports

As for the perceived technological infrastructure and supports in participants' work environment (Figure 6-4), approximately half of the participants rated "sufficient" on tools/resources for finding literature and managing citations. Similarly, for tools supporting other data production activities such as collecting, processing and analyzing data, the rating of "insufficient" is less than 12%.

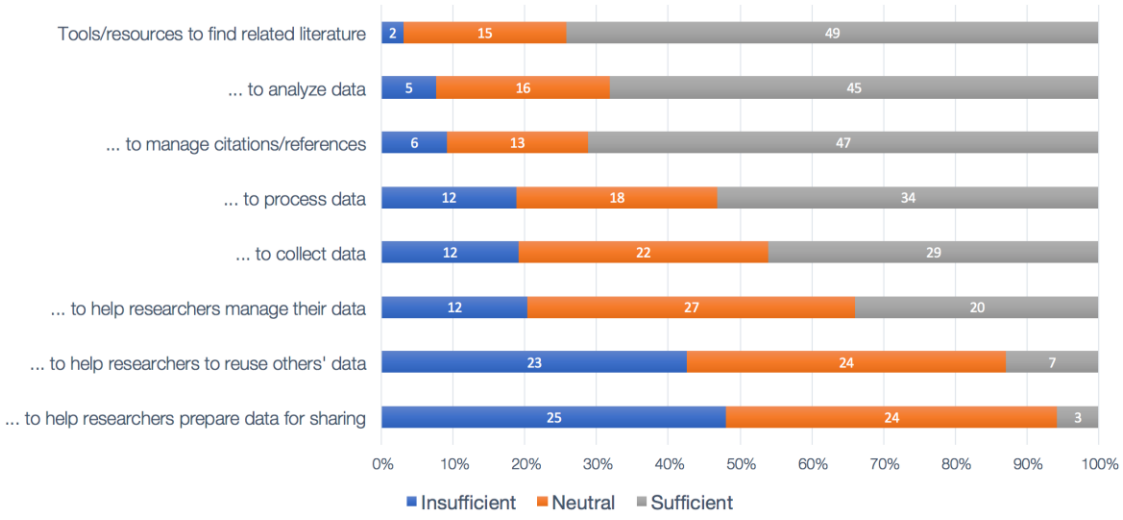


Figure 6-4. Technological supports

On the contrary, only a small portion of participants report that tools or resources for facilitating data reuse (n=7, 13%) and data sharing (n=3, 5.8%) are sufficient, suggesting that the related research data services have room for improvement to prepare social scientists for data sharing and reuse. Participants were further asked to identify the persons involved in the research data services or supports in their institutions in a multiple selection question. For both PITT and CMU (Table 6-4), the majority of participants selected libraries and researchers' own colleagues as supporters.

Table 6-4. Internal human resource supports in work environment

Human resources	PITT	CMU
Research support unit(s)*	34%	18.80%
The university's library systems	80%	50%
The university's IT support unit(s)	40%	6.30%
Administrative office(s)	12%	25%
Designated data manager(s)	12%	6.30%
Colleagues	74%	81.30%

Note: each question is preceded by a context description: "Based on your past impressions, which of the following are involved in these research data services in your work environment?"
*e.g., Office of Research at Pitt; Office of Sponsored Programs at CMU

Table 6-5. Perceived benefits

Perceived benefits	M	SD	1	2	3	4	5
More citations	3.38	.890	1.5%	10.6%	48.5%	27.3%	12.1%
Career advancement	3.32	.931	3.0%	13.6%	40.9%	33.3%	9.1%
Collaboration opportunity	4.08	.771	1.5%	3.0%	7.6%	62.1%	25.8%
Fulfill others' research need	3.94	.892	0%	3.0%	33.3%	30.3%	33.3%
Inspire other researchers	4.11	.767	0%	1.5%	19.7%	45.5%	33.3%

Note: each question is preceded by "The following statements relate to your thoughts about sharing data with others. Please tell us how much you agree with the following statements."

6.3.6 Individual motivations

The participants were also asked about the perceived benefits of and rewards for sharing data, as reported in Table 6-5 (1: strongly disagree; 5: strongly agree). More than 85% of participants (strongly or slightly) agree that opportunity for collaboration is a benefit of data sharing. However, it is interesting that a large percentage of participants (more than 40%) take a neutral stance regarding citations and career advancement.

It is worth noting that two of the perceived benefits (i.e., Fulfill others' research need and Inspire researchers outside your field) are altruistic. If considering only the "strongly agree" column, these two altruistic reasons outperform the rest, and they are each backed by 33.3% of participants.

6.4 SUMMARY OF CASE STUDY 1

CS1 presents a profile instrument that captures individual social scientists' research activities, data-sharing practices, data characteristics, and perceived technological support. In this case study, research activities and data-sharing practices in three participant groups are investigated, and there are no significant differences among social scientists who prefer quantitative, mixed, and qualitative methods. This result may imply that researchers with different research methods may share similar contexts, barriers, or drivers.

The results confirm that early-career social scientists rarely share data, which is largely consistent with prior work, as well as the observations in PS1 and PS2. However, as a baseline, manuscript sharing in social sciences is not much more frequent than data sharing. Scholarly altruism is also found to be a common reason to share data, whereas extrinsic motivations (e.g., gaining citations and career advancement) are less relevant in this case study.

Most importantly, a chasm is revealed between early-career social scientists' attitudes, beliefs, and actual behaviors: social-science researchers highly value data sharing and witness data sharing in their fields, but they do not actually share their own data. This observation is consistent with Preliminary Study 1.

Case Study 1 benefits the overall design and outcome of this dissertation study. Specially, the implications of low data-sharing frequency in CS1 are two-fold. First, it is imperative to include more

participants with data-sharing experience in the next stage. This implication strengthens the importance of the sampling method in Case Study 2. Second, there is a critical need to not only study motivations and incentives, but also the "barriers" in the way of social scientists' data sharing. This implication inspired the design of the focus group protocol for Case Study 3.

7.0 CASE STUDY 2: QUALITATIVE DATA SHARING PRACTICES IN SOCIAL SCIENCES

7.1 OVERVIEW OF CASE STUDY 2

The results described in the previous chapter (Case Study 1) show that early-career researchers do not have much experience sharing data at discipline data repositories. To obtain enough research samples, this case study targets people with experience sharing qualitative data; that is, people who have previously shared data or have been involved in a study that has deposited data in a data repository.

This case study aims to 1) present descriptive statistics and describe the knowledge infrastructure of qualitative data sharing, and 2) further examine factors that influence social scientists' data-sharing behaviors, such as perceived technologies, extrinsic motivations, and intrinsic motivations.

This case study plays two important roles in this dissertation study. First, it acts as a refined version of Case Study 1 by considering a more representative sample, as well as including more specific questions regarding qualitative data. Second, this study complements CS1. Since CS2 mostly comprises senior researchers involved in mixed-method or qualitative studies, it can be used to triangulate the perceptions of early-career social scientists' in CS1.

7.2 RESEARCH SITES

To achieve the study goal, CS2 samples researchers who have the following experience:

- Those who have shared data in data repositories in the past ten years (2006-2015, and the first four months in 2016);
- Those who have shared qualitative data

For the first consideration, potential participants are targeted in two data repositories: the Interuniversity Consortium for Political and Social Research (ICPSR) and the Qualitative Data Repository (QDR).

Interuniversity Consortium for Political and Social Research (ICPSR). The Interuniversity Consortium for Political and Social Research (ICPSR) was established in 1962 and is the world's largest primary data archive of social science research. As of July 2016, ICPSR holds 8,053 studies, 68,033 datasets, and 196,881 files for download (ICPSR, 2016).

Although ICPSR is the oldest and most representative data repository in social science, qualitative data is not sufficiently represented. An additional repository, the Qualitative Data Repository (QDR), is selected to fill this gap.

Qualitative Data Repository (QDR). QDR is a qualitative data repository hosted by the Center for Qualitative and Multi-Method Inquiry, under the Maxwell School of Citizenship and Public Affairs at Syracuse University. QDR was founded in 2013 and hosts 27 research projects; it also offers a variety of resources or guidance related to sharing qualitative data. However, QDR has just started and as of Summer 2016, has only 35 PIs who can be reached on the website.

7.3 DATA COLLECTION

7.3.1 Sampling

To take advantage of both data repositories, the sample includes all the PIs in QDR (which contains a small number of researchers who have all shared qualitative data), as well as ICPSR PIs who might have deposited qualitative data. To achieve this, CS2 took advantage of the dataset keywords on ICPSR and identified potential PIs by performing relevant keyword searches with a ten-year span. This ten-year timeframe ensures that the collection of PIs reflects their most recent status. Table 7-1 summarizes a possible candidate list of keywords based on ICPSR's suggested "examples of types of qualitative data that may be archived for secondary analysis" (ICPSR, 2012, p.27): *interview*, *qualitative analysis*, *qualitative study*, *focus group*, and *field study*, and the number of studies each keyword returns on ICPSR.

Table 7-1. A set of search keywords as of April 17, 2016

Keywords	Results (all time)	2006-2016 (ten-year span)
Survey	2719	399
Questionnaire	1818	302
Interview	2081	306
Qualitative analysis	373	109
Qualitative study	388	111
Qualitative method	319	88
Qualitative research	390	111
Focus group	1375	279
Field study	1823	320
Field trip	443	101
Mixed method	955	198
Historical research	1016	145
Historical study	980	145
Oral history	455	115
Case study	2782	415
Press clipping	123	31
Delphi	16	2
ICPSR's "qualitative" tag	65	37

After using the results returned by these keywords and removing any duplicates (for example, returned studies containing both keywords "Qualitative method" and "Qualitative study" are highly overlapped), 1,272 study identifiers are identified (hereafter: study IDs). The metadata (Dublin Core) of a study on ICPSR can be accessed via a URL that comprises an address of "icpsr.umich.edu/icpsrweb/neutral/dc/studies/" followed by its study ID. For instance, the public URL of metadata for the study "Prescription for Health Evaluation: Practice Information Form Data, 2005-2007" (study ID 27041) can be accessed at "icpsr.umich.edu/icpsrweb/neutral/dc/studies/27041." The metadata of 1,272 studies was accessed and saved as an XML file format by a Python script. Another Python script was used to extract the "creator" field and parse data points into a CSV file. Among the 1,272 studies, 909 valid names are extracted. Among these, 744 researcher profiles, CVs, or personal websites were found online via search engines or academic social media networks such as ResearchGate, and 694 emails are manually collected. Figure 7-1 summarizes the sampling and data collection process in Case Study 2.

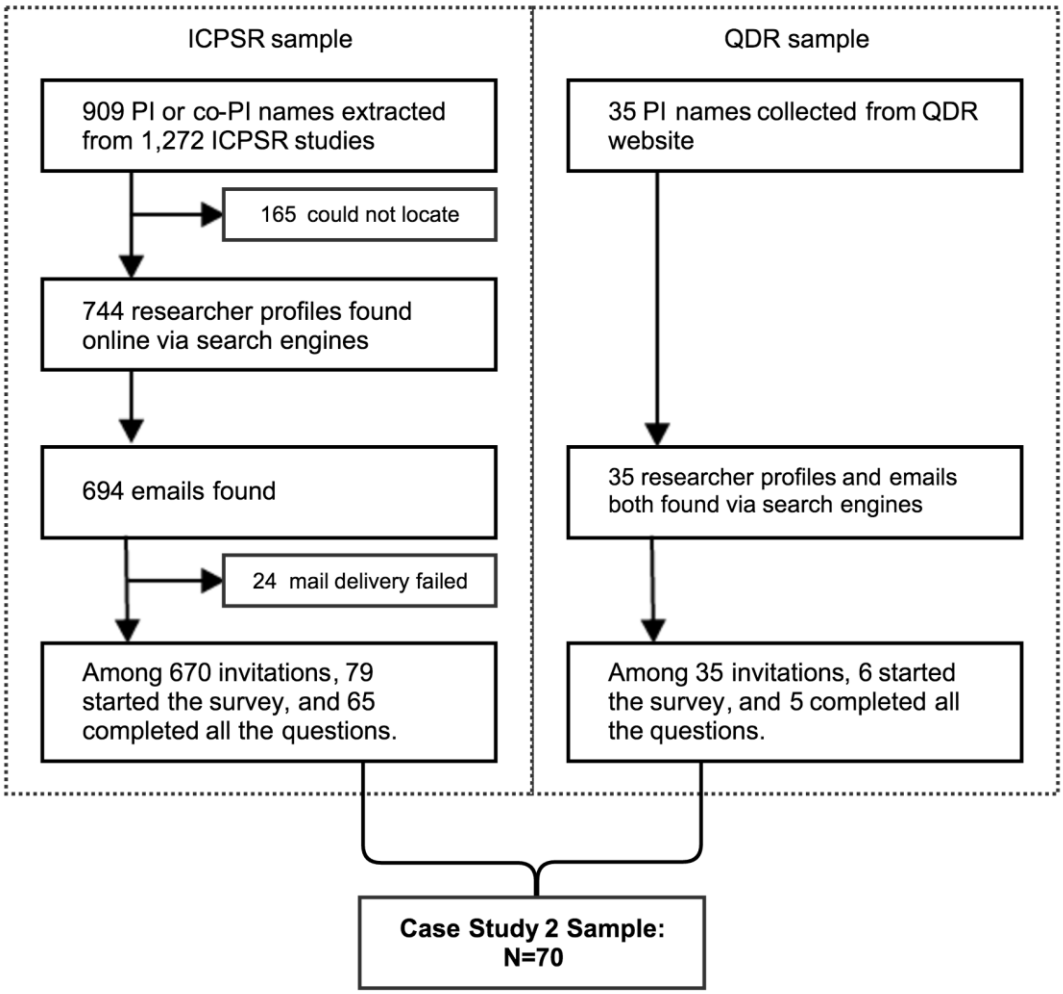


Figure 7-1. Overview of sampling and responses

7.3.2 Survey distribution

In August 2016, the questionnaire invitations were sent (via Qualtrics software program) to these 694 potential participants. Among them, 24 invitation were immediately rejected and displayed as "bounced" on the Qualtrics email management system due to delivery failure. Among the remaining valid invitations, 79 participants started the survey, resulting in a response rate of 11.8%. Sixty-five out of the 79 participants completed the survey and are included in the sample.

The same procedure was repeated for retrieving the contact information from QDR. Thirty-five emails were found. After sending out invitations, six participants started the survey, yielding a response rate of 17.1%, and five out of six were completed. The final sample of CS2 is N=70.

7.3.3 Demographics of participants

The analyses in this section are based on the completed responses of 70 participants. Most participants come from higher education: 84.5% of participants report their work sector as "academic" and 41.4% are full-rank professors. The top three age groups that participants report are 45-54 (42.9%), 55-64 (20%), and 35-44 (18.6%). Forty-four participants are male (62.9%). The detailed demographic statistics are reported in Data Table 1 in APPENDIX F.

In this case study, disciplines are categorized into ten groups. The discipline options are based on NSF's Survey of Earned Doctorates as shown in Data Table 2 in Appendix F. The top five discipline groups are 1) political, government & policy (n=16), 2) law, criminology & criminal justice (n=12), 3) sociology & social work (n=11), 4) public health & family studies (n=11), and 5) psychology & decision science (n=9). The results from the top disciplines (such as political science and psychology groups) are similar to Case Study 1, in which the data were collected from all social

sciences units from the University of Pittsburgh and Carnegie Mellon University. Disciplines such as education and information & communication studies are less represented in Case Study 2. On the other hand, a significant portion of participants are from criminology and public health fields. The reason for this might be that ICPSR features the National Archive of Criminal Justice Data (NACJD).

Table 7-2. Distribution of discipline groups in Case Study 2

Discipline groups	N	%
Economics & Business	3	4.3
Education	1	1.4
Geography	1	1.4
Info and Communication	2	2.9
Law, Criminology & Criminal Justice	12	17.1
Political, Government & Policy	16	22.9
Psychology & decision making	9	12.9
Public health & Family	11	15.7
Sociology & Social Work	11	15.7
Social Sciences, General	4	5.7

The word cloud in Figure 7-2 presents the research interests of the participants, collected from their open-ended responses. Frequent keywords include health, violence, family, elections, and comparative politics.

Figure 7-2. Word cloud of research interests in Case Study 2

Participants are asked to estimate the proportion of their data in a five-point scale. Specifically, the question is: "Please recall one of your most recent research projects and estimate the proportion of your qualitative (QUAL) data, compared with your quantitative (QUANT) data in it." To ensure that every participant has an identical definition of *qualitative*, the following statement precedes the question:

"*Qualitative data are data generated from qualitative approaches or involve qualitative judgments, such as interviews, open-ended surveys, focus groups, oral histories, observations, or content analysis.*"

The results are summarized as follows.

Although CS2 targeted people who might have experience with qualitative data via a keyword search in ICPSR, the final sample still contains 22 participants (31.0%) who claim that they are doing purely quantitative studies. As shown in Table 7-3, among those who report some portion of qualitative data, most use mixed methods and are leaning toward quantitative. That is, few scholars conduct purely quantitative studies (the same finding as in Case Study 1). Understanding the

cause of such a phenomenon can be interesting (i.e., qualitative scholars' self-identification) but is beyond the scope of this dissertation work.

Table 7-3. Qualitative data proportion (N=70)

Proportion	N	%
Purely qualitative	3	4.2%
Mixed but more qualitative	5	7.0%
Equal mix of qualitative and quanitative	11	15.5%
Mixed but more quantitative	30	42.3%
Purely quantitative	22	31.0%

There are some contradicting responses between participants' experienced data type and their data proportions. Specifically, four participants report that they have experience with qualitative data (e.g., data generated from field observations), but still claim they are doing purely quantitative research. Such contradicting responses might arise from how the questions were phrased in Instrument 2: the participants were first asked to state the type (source) of data they interact with in their research career, and then to estimate the proportion of qualitative data in their *most recent* research project(s). Therefore, these participants might have dealt with qualitative data before but rarely do so anymore, or they are always conducting purely quantitative studies but have collaborated on qualitative projects in the past.

7.4 DESCRIPTIVE RESULTS

7.4.1 Data characteristics

The social scientist participants in this case study report that their most common data source comes from informants, such as direct responses from surveys, interviews, or focus groups (Figure 7-3); 92.9% of participants have experience with this type of data. The second and third most common data sources are secondary data (77.1%) and observational data (45.7%), respectively. This ranking order is largely similar to that of CS1, but the percentage of each data source in CS2 is slightly higher. This is self-explanatory because CS2 comprises more senior faculty who has spent more time in the academics.

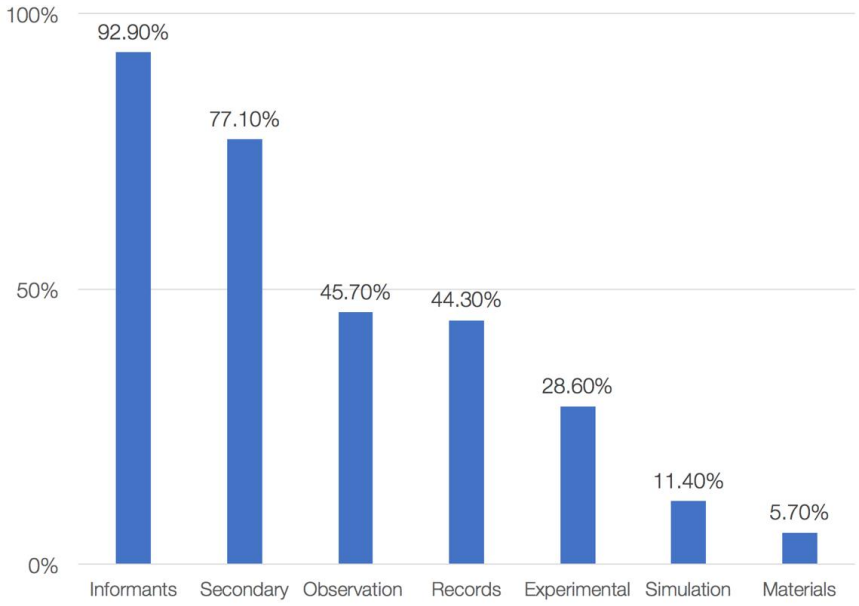


Figure 7-3. Most common data type (source)

The data sources seem very diverse across all disciplines in Case Study 2 in Figure 7-4, unlike Case Study 1 in which only a few discipline groups have diverse sources of data. The reason for this difference might be that senior faculty stays in academia longer and thus has dealt with more projects and diverse data sources.

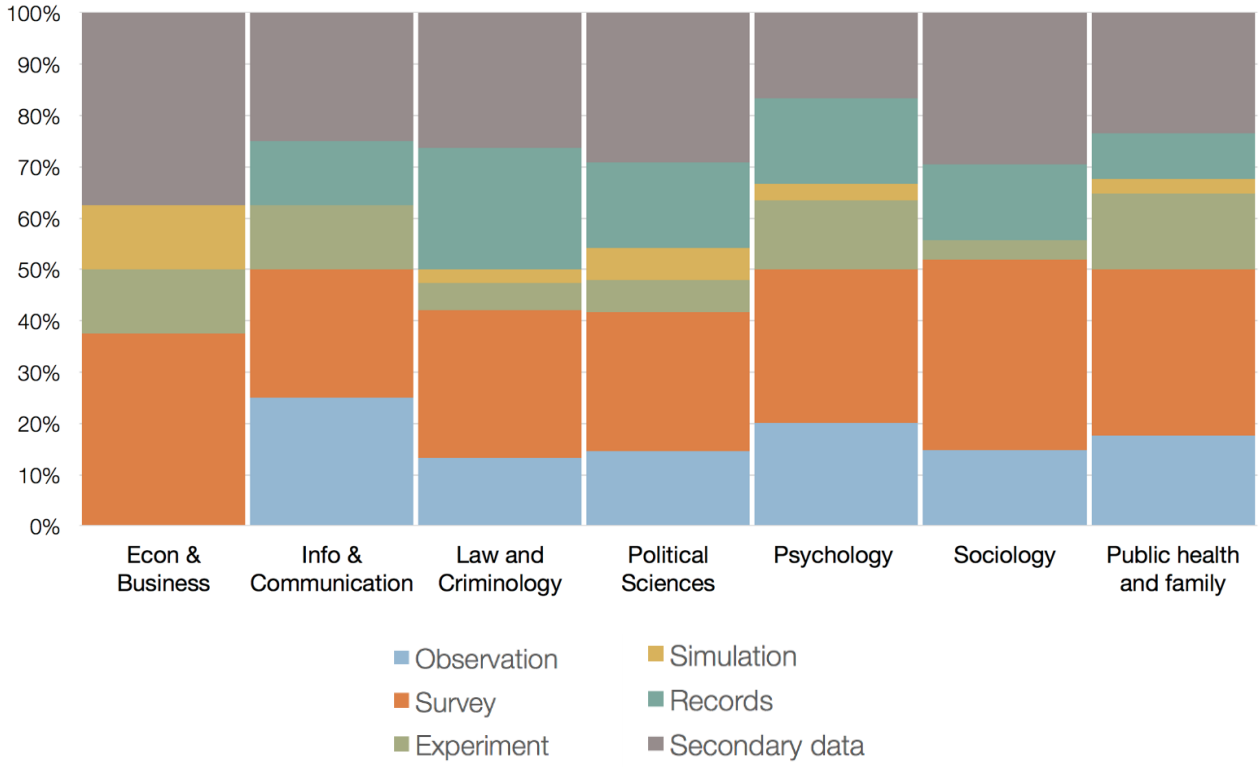


Figure 7-4. Data types and disciplines in Case Study 2

Note: education and geography is omitted.

To further examine what kind of qualitative data is sharable, a subset of 48 social scientists was selected by excluding the "purely quantitative (n=22)" participants from the dataset.

As shown in Table 7-4, a total of 98% (i.e., 91.7% very likely and 6.3% somewhat likely) of participants unanimously agree they are likely to share the detailed procedures of data collection (e.g., interview protocols). Also, 87.3% of participants are likely to share "survey instruments with actual question items" and more than half are likely to share "analytic scripts" (16.7% somewhat likely and 45.2% very likely).

As for the survey responses (with individual responses), the result is polarized: 34.8% are strongly unlikely whereas 41.3% are strongly likely to share. It is surprising that about 50% of participants report they are very likely to share interview transcripts. The responses on research notes show a lack of agreement: every option is evenly distributed, ranging from 10% to 30%. The results suggest that sharing procedures, instruments, and analytic scripts receive the most collective agreement, which can be an important reference for developing the best qualitative data-sharing practices.

Table 7-4. Shareable data deemed by participants (n=48)

Types of qualitative data	n	Mean	SD	1	2	3	4	5
Detailed procedure of data collection (e.g., interview protocol)	48	4.85	.62	2.1%	0%	0%	6.3%	91.7%
Survey instrument with actual question items	47	4.57	1.02	4.3%	2.1%	6.4%	6.4%	80.9%
Analytic scripts	42	3.69	1.49	14.3%	9.5%	14.3%	16.7%	45.2%
Multi-media	23	3.52	1.47	17.4%	4.3%	21.7%	21.7%	34.8%
Survey response (with individual responses)	46	3.22	1.81	34.8%	6.5%	2.2%	15.2%	41.3%
Interview transcripts	43	3.05	1.53	25.6%	14.0%	11.6%	27.9%	20.9%
Researcher notes	45	3.04	1.50	20.0%	22.2%	15.6%	17.8%	24.4%

Note: each item is preceded by "Based on your overall experience, which data or materials at below would you be willing to share with other researchers? 1: Very unlikely; 2: Somewhat unlikely; 3: Neutral; 4: Somewhat likely; 5: Very likely"

7.4.2 Perceived technologies

This section reports the technological infrastructure as well as technological supports that are perceived by the participants in their work environment.

Table 7-5. Descriptive statistics of technological supports in Case Study 2

Attributes	M	SD	1	2	3	4	5
analyzing data	4.34	0.866	0.0%	4.3%	12.9%	27.1%	55.7%
collecting data	3.84	1.072	1.4%	11.4%	22.9%	30.0%	34.3%
discovering others' data	3.13	1.115	5.7%	28.6%	22.9%	32.9%	10.0%
preparing data for sharing	2.66	1.25	21.4%	25.7%	28.6%	14.3%	10.0%

Note: each item is preceded by "In my work environment, technology related to...; 1: Very insufficient; 2: Somewhat insufficient; 3: Moderate; 4: Somewhat sufficient; 5: Very sufficient"

This section reports the technological infrastructure as well as technological supports that are perceived by the participants in their work environment. This section reports the technological infrastructure as well as technological supports that are perceived by the participants in their work environment.

Table 7-5 demonstrates that the perception of supports and tools for data discovery (i.e., finding data for reuse) and sharing are both rated least sufficient among tools for supporting data production (analyzing data and collecting data).

Figure 7-5 displays the comparison of three variables between Case Study 1 (early-career social scientists, marked in blue circles) and Case Study 2 (marked in orange diamonds). Since the original instrument in CS1 is designed to be exploratory, the scale is only 1 (insufficient), 2 (moderate), and 3 (sufficient). In order to compare these two cases, the 1-5 scale in CS2 was recoded as 1 (insufficient), 2 (moderate), and 3 (sufficient). A Mann-Whitney test suggests that there are

significant differences at the .05 level between the two case study samples in terms of technological supports in data analysis (U = 1972, p = .049) and technological supports in data collection (U = 1805, p = .044). Both mean ranks in CS2 were higher than those in CS1. Technological supports for preparing data for sharing in CS2 have a higher rating on average, but there is no statistical significance found in their distribution.

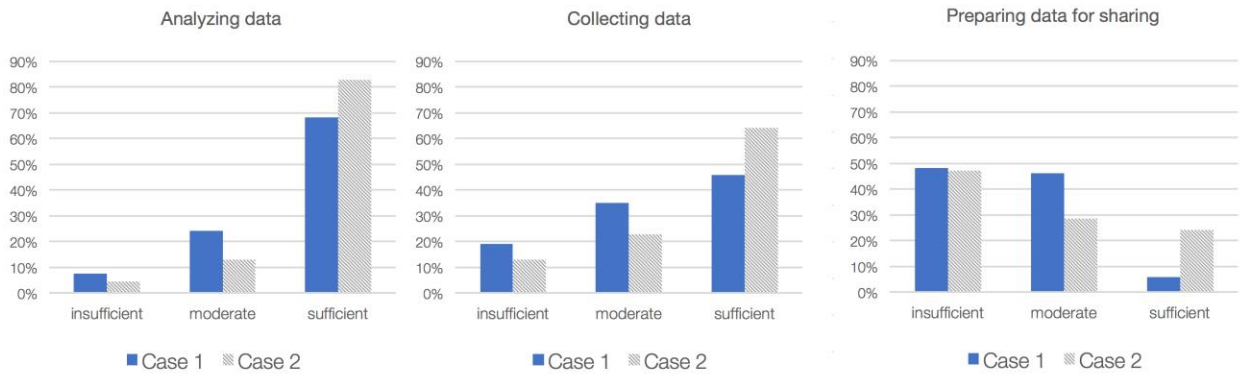


Figure 7-5. Distributions on technological supports in two studies

7.4.3 Perceived discipline community culture

Like CS1, CS2 also examines community culture regarding qualitative data sharing.

Table 7-6 shows a list of possible community cultures and to what extent the participants agree that these are the community cultures. Note that 1 represents strongly disagree and 5 represents strongly agree.

Table 7-6. Descriptive statistics of discipline community culture in Case Study 2

Community culture	M	SD	1	2	3	4	5
Common to see people sharing their data.	3.23	1.206	8.6%	24.3%	15.7%	38.6%	12.9%
There is a generic standard for data sharing.	2.73	1.35	22.9%	25.7%	20.0%	18.6%	12.9%
People care a great deal about data sharing.	3.44	1.163	5.7%	17.1%	24.3%	32.7%	20.0%

Note: each item is preceded by "To what degree do you agree with the following statements describing your discipline community in terms of data sharing? In my discipline community…; 1: Strongly disagree; 2: Somewhat disagree; 3: Neither disagree or agree; 4: Somewhat agree; 5: Strongly agree"

The majority of participants (strongly or slightly) disagree with the existence of a standard procedure or a well-known, recognized data infrastructure. The result is consistent with Preliminary Study 1's (i.e., Jeng & Lyon, 2016) findings that standards are one of the least-developed capabilities in social science disciplines. By comparing Case Studies 1 and 2 in terms of the responses for "there is a data sharing standard" and "people [in the discipline community] care a great deal about data

sharing", it can be seen that more participants give a high rating in CS2 than in CS1, as shown in Figure 7-6. The early-career social scientists in CS1 seem to disagree with the statement that the community cares a great deal about data sharing, whereas CS2 participants have higher ratings.

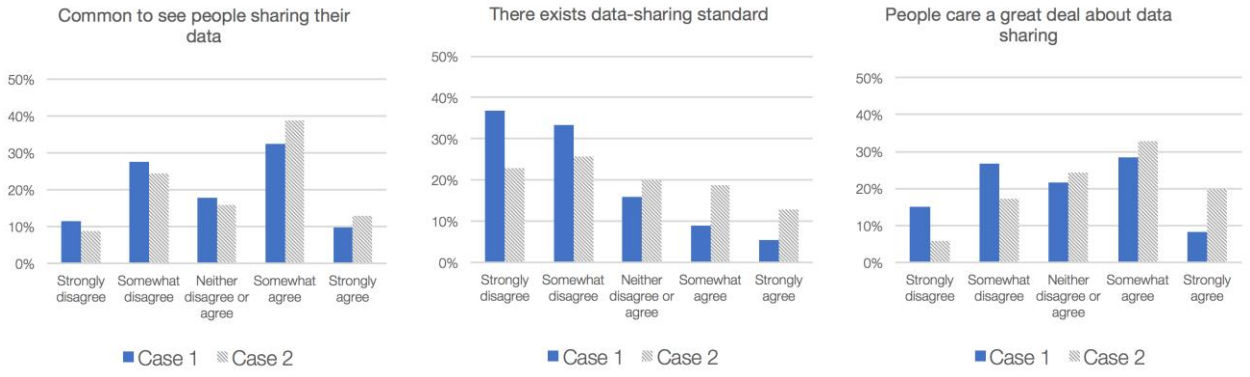


Figure 7-6. Distributions on discipline community culture in two studies

Consistent with the observation in Figure 7-6, a Mann-Whitney U test suggests there are significant differences at the 0.05 level between the two case study participants' perceptions regarding "there is a data-sharing standard" ($U = 1418.5$, $p = 0.009$) and "people [in the discipline community] care a great deal about data sharing" ($U = 1568$, $p = 0.011$). These U test results show that CS2 participants have a higher rating on these two variables on average.

7.4.4 Individual motivation and concerns

Participants in CS2 were similarly asked about motivations (Table 7-7) for data sharing, as reported in the following tables. Again, the score ranges from 1 to 5, with 1 representing strongly disagree and 5 representing strongly agree.

Table 7-7. Descriptive statistics of individual motivations in Case Study 2

Individual motivations		M	SD	1	2	3	4	5
Intrinsic motivations	Inspire other researchers	4.19	.873	1.4%	2.9%	12.9%	41.4%	41.4%
	Help others to fulfill their research needs	4.53	.737	0%	2.9%	5.7%	27.1%	64.3%
	Sample to impart the social research method	4.13	.883	1.4%	4.3%	11.4%	45.7%	37.1%
Extrinsic motivations	Collaborate with others	4.03	.884	0%	7.1%	15.7%	44.3%	32.9%
	More citations	3.71	1.08	2.9%	10.0%	28.6%	30.0%	28.6%
	Career Advance	3.61	1.07	4.3%	8.6%	31.4%	32.9%	22.9%

Note: each item is preceded by "The following statements relate to your thoughts about sharing data with others. Please tell us how much you agree with the following statements. Data sharing can...; 1: Strongly disagree; 2: Somewhat disagree; 3: Neither disagree or agree; 4: Somewhat agree; 5: Strongly agree"

While intrinsic motivations have the highest averages, more than half of the participants strongly agree or somewhat agree with the statement that data sharing can help collaboration with others, increase citations and advance careers.

Compared with CS1 participants, CS2 participants are found to have significantly higher ratings on "help others to fulfill their research needs" ($U = 1445$, $p < 0.00001$) and "gaining more citations" ($U = 1838.5$, $p = 0.031$). That is, the senior social scientists in CS2 concur with the statement that data sharing can fulfill others' research needs. Moreover, the statistically significant difference in "more citations" also shows that the senior social scientists in CS2 tend to agree that an increase of

citations is a motivation to share data. The distribution of ratings is shown in Figure 7-7 and Figure 7-8.

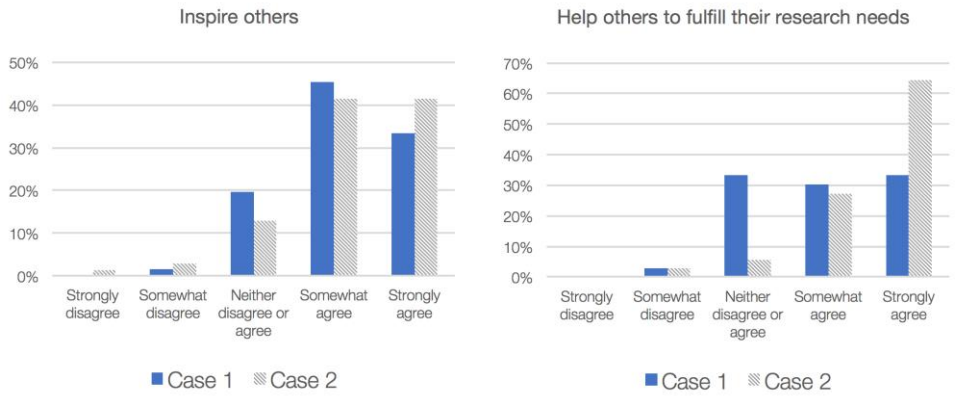


Figure 7-7. Distributions on intrinsic motivations in two studies

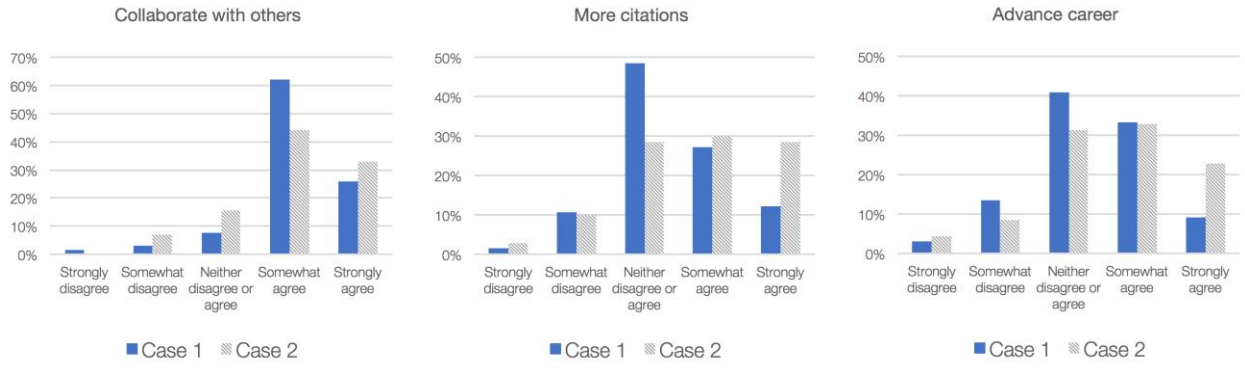


Figure 7-8. Distributions on extrinsic motivations in two studies

The open-ended responses provided by the survey participants in CS2 also reveal different levels of concern about data sharing in social sciences. Twenty-two out of 70 participants (31.4%) left comments or suggestions at the end of the questionnaire, many of which are related to data-sharing factors.

Two main messages stand out. First, the participants repeatedly stress that ethical considerations are the most critical in terms of sharing data in social sciences:

> *"whether to share data, and what data, is the risk to human subjects. It can be a major obstacle to data sharing" (P93, or P10 in CS2).*

Another participant mentioned that confidentiality concerns and disclosure risks are "huge issues":

> *"Confidentiality and deductive disclosure are huge issues for me re: data sharing, since all of my research is about risk behaviors ([e.g.] sexual violence[2]) and much of it involves minors" (P86, P29 in CS2).*

Second, according to several participants, funder pressure is the most critical factor in data sharing. One participant mentioned that he works on an NIA-funded project (i.e., National Institute on Aging) and is required to share data:

> *"I work on a NIA-funded study…I HAVE to share my data and it doesn't matt[e]r if I have enough time, money, etc. to do so" (P128, or P22 in CS2).*

---

[2] Mentioned topics are converted to a more general interest for the participant's identity protection.

Another participant (P73, P54 in CS2) describes the tension she faces between funders' requirements and the concerns about confidentiality:

> *"I have only deposited data because it was required by federal grants, and even then was hesitant due to confidentiality concerns" (P73, P54 in CS2).*

7.4.5 Data sharing practices

This section discusses the descriptive results of data-sharing behaviors among Case Study 2 participants. Table 7-8 reports the participants' responses on different data-sharing channels. A higher score means a higher level of involvement in qualitative research, where 1 represents Never or Rarely and 5 represents Frequently or Always. Every participant was shown this scale:

1. Never or Rarely (about 0-10% of the time)
2. Occasionally (about 25% of the time)
3. Sometimes (about 50% of the time)
4. Often (about 75% of the time)
5. Frequently or Always (about 90-100% of the time)

Table 7-8. Data sharing behaviors and participants' preferred methods

	purely quant (22)		QUANT more (30)		Equal (10)		QUAL more (5)		Purely QUAL (3)	
	M	SD	M	SD	M	SD	M	SD	M	SD
Institution repository	2.77	1.510	3.37	1.542	3.00	1.491	3.00	1.581	2.00	1.732
Public Web spaces	2.5	1.626	2.1	1.398	2.2	1.549	1.6	1.342	1	0
Academic social media	1.45	1.184	1.7	1.291	1.7	0.949	1	0	1	0
Discipline data repositories	3.05	1.495	3.33	1.348	3.5	1.354	3.4	1.517	1.67	0.577
Via emails	2.59	1.563	2.73	1.363	3.1	1.287	3.4	1.517	1.67	0.577
Publications as supplemental materials	2.23	1.51	2.27	1.437	2.1	1.101	2	1.414	2	1

Note: 1. Never or Rarely (about 0-10% of the time); 2. Occasionally (about 25% of the time); 3. Sometimes (about 50% of the time); 4. Often (about 75% of the time); 5. Frequently or Always (about 90-100% of the time)

An interesting observation is that participants who are involved in mixed methods (QUANT more, Equal, and QUAL more) report a higher frequency of official channels such as "Institution repository" and "Discipline data repositories." This contradicts the common-sense assumption that quantitative researchers are more likely to share data. Moreover, in sharing via email (upon request), QUAL More was rated higher than pure QUANT.

All data were ranked before a Kruskal-Wallis H test (χ^2), the results of which suggest no statistical difference across three categories of proportion of qualitative data (none, partial, and more than half) in terms of job characteristics. That is, these observed differences were not statistically significant.

7.5 FACTORS INFLUENCING QUALITATIVE DATA SHARING

In this section, CS2 further examines factors that influence data-sharing practices of social scientists who have recently dealt with qualitative data (n=48; participants who answered "purely quantitative

[n=22]" were excluded). In practice, researchers have suggested there should be at least 10 (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996) to 15 (Babyak, 2004) incidents per predictor (a.k.a., event per variable, EPV). This sample subset is legitimate to run a multiple linear regression with four predictors.

7.5.1 Hypothesis development

Continuing with the instrument refinement in Chapter 5, independent variables (i.e., possible predictors) are listed in Table 7-9. Cronbach's alpha is used to measure agreement and consensus among different items in each variable. As mentioned in Chapter 5, the acceptable values of alpha should be equal or above 0.70 (Gliem, J. & Gliem, R., 2003), because this ensures that the internal consistency in these seven variables is acceptable or good. The list of hypotheses developed for each independent variable are listed in Table 7-10. The dependent variable (i.e., the outcome being predicted) is the sharing behavior.

Table 7-9. The reliability of independent variables

Independent Variables	Number of items	Cronbach's alpha
Trust of data quality and that it will be reused	4	.841
Intrinsic motivations	3	.852
Extrinsic motivations	3	.782
Ease of sharing	3	.782
Tech supports	4	.800
Discipline community practice	3	.725
Data ownership	2	.821

Table 7-10. Hypothesis of data sharing behaviors

Themes	Hypotheses
Individual motivations	H1: Perceived extrinsic benefits would positively influence data sharing behaviors
	H2: Perceived intrinsic benefits would positively influence data sharing behaviors
	H3: Perceived ease of sharing would positively influence data sharing behaviors
Data ownership	H4: Perceived data ownership would positively influence data sharing behaviors
	H5: Perceived trust of data quality and that it will be reused would positively influence data sharing behaviors
Community	H6: Perceived community practice on data sharing would positively influence data sharing behaviors
Technology	H7: Perceived technological support would positively influence data sharing behaviors

Table 7-11 summarizes the correlation results of each factor after creating the subset of 48 participants. Among these factors, 1) perceived intrinsic motivations (IM for intrinsic motivations), 2) perceived extrinsic motivations (EM for extrinsic motivations), and 3) perceived technological support (TS for technological support) have significant positive correlation with social scientists' data-sharing frequency within the past three years. Discipline community culture (DC for discipline community) is not found to have a correlation with data-sharing frequency. Figure 7-9 illustrates the scatter plots to help determine correlation.

Table 7-11. Correlation table

	DF	EM	IM	ES	DO	TD	TS	DC
Data-sharing frequency in past three years **(DF)**								
Extrinsic motivations **(EM)**	.427**							
Intrinsic motivations **(IM)**	.390**	.529**						
Ease of data sharing **(ES)**	.346*	0.233	0.174					
Data ownership **(DO)**	0.096	0.106	0.003	0.244				
Trust of data quality and that it will be reused **(TD)**	0.204	0.184	-0.039	.502**	0.193			
Perceived technological support **(TS)**	.402**	0.237	.471**	0.191	-0.08	0.256		
Discipline community culture on data sharing **(DC)**	0.225	.348*	0.239	-0.033	.313*	0.12	0.191	

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

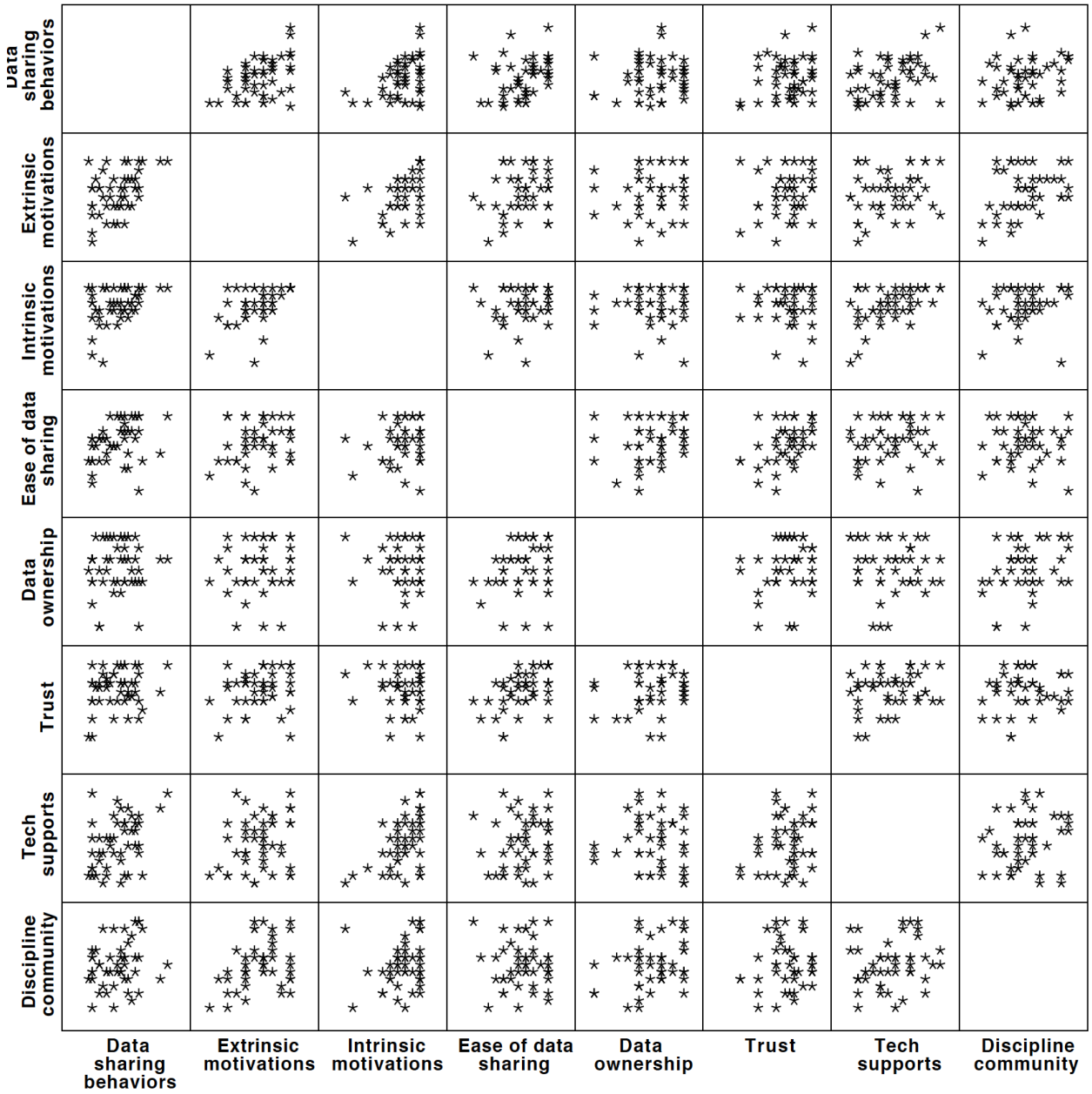


Figure 7-9. Scatter plots of correlated variables based on Table 7-11

7.5.2 Linearity

A multiple linear regression was undertaken to examine the variance in social scientists who have experience sharing data frequently. The independent variables *trust*, *ownership*, and *discipline culture* have been excluded based on the correlation result.

The histogram of the residuals in Figure 7-10 looks symmetric and fairly unimodal, which illustrates an approximately normal distribution of residuals. P-P (probability–probability) plots are used to evaluate the skewness of a distribution. The plot will approximately present as a linear shape when the specified theoretical distribution is the correct model. The normal probability plot in Figure 7-11 looks more or less linear. Both Figure 7-10 and Figure 7-11 show that the deviation is fairly normally distributed.

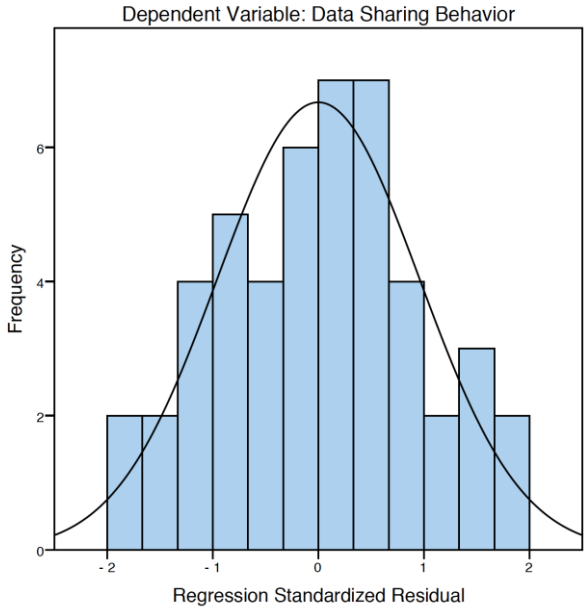


Figure 7-10. Histogram of standard residual

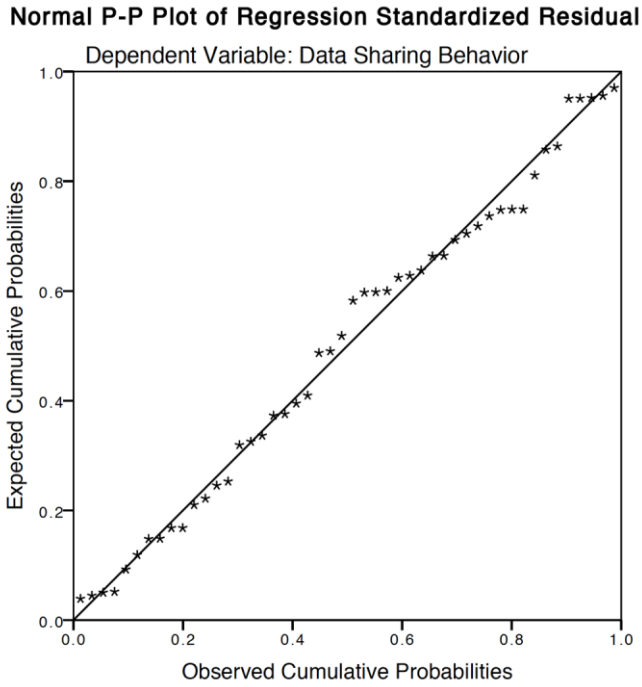


Figure 7-11. The normal P-P plot of regression standardized residual

The model was calculated to predict data-sharing frequency based on the above-mentioned four possible variables using the Stepwise method:

- Model 1: Enter variable EM- perceived extrinsic motivation as the only independent variable
- Model 2: Enter variable TS- perceived technology support into Model 1
- Variables IM-perceived intrinsic motivation and ES-ease of data sharing were excluded in both Model 1 and Model 2.

Table 7-12 lists two models for consideration. After evaluating by the F and the coverage of R^2, Model 2 is selected. The R square value (0.278) in Model 2 represents the scattered points around the regression line. This explains a significant model, $F(2, 45) = 8.669$, $p = 0.0006$), that predicts 27.8% of the sample outcome variance. The R square value here is comparable to related work (e.g., 28%-39% in Curty, 2016; 18.4% in Kim and Alder, 2015). The tolerance and variance

inflation factors (VIF) are diagnostic factors that help identify multicollinearity. The tolerance of collinearity in both models ranges from 0.944 to 1.0; the VIFs are satisfactory (<2.5), ensuring no multicollinearity. Table 7-13 presents the summary of the hypothesis results.

Table 7-12. Models

Predictor variable	R	R^2	F	P	t	p	Collinearity tolerance	VIF
Model 1	.427	.183	10.284	.002**				
Extrinsic motivations					3.207	.002**	1.000	1.000
Model 2	.527	.278	8.669	.0006***				
Extrinsic motivations					2.700	.010*	.944	1.06
Technological support					2.439	.019*	.944	1.06

Note: ***: p<.001, **: p<.005, *: p<.05
The predictor perceived extrinsic motivation was entered into the Model 1, B_{em} = 0.948, t = 3.207, p=0.002). Model 2 is based on Model 1, and the perceived technological support is entered, resulting in perceived extrinsic motivation (B_{em} = 0.781, t = 2.7, p=0.01) and perceived technological support with B_{ts} = 0.494, t = 2.439, p=0.019.

Table 7-13. Summary of hypothesis results

Themes	Hypothesis	Results
Individual motivations	H1: Perceived extrinsic benefits would positively influence data sharing behaviors	Supported
	H2: Perceived intrinsic benefits would positively influence data sharing behaviors	Not supported
	H3: Perceived ease of sharing would positively influence data sharing behaviors	Not supported
Data ownership	H4: Perceived data ownership would positively influence data sharing behaviors	Not supported
	H5: Perceived trust of data quality and that it will be reused would positively influence data sharing behaviors	Not supported
Community	H6: Perceived community practice on data sharing would positively influence data sharing behaviors	Not supported
Technology	H7: Perceived technological support would positively influence data sharing behaviors	Supported

Note: who has self-identified as mixed- or qualitative-preferred researchers

7.6 SUMMARY OF CASE STUDY 2

The findings in Case Study 2 can be highlighted as follows:

- Participants (who have shared qualitative data in ICPSR and QDR) are more likely to share research products related to methodological aspects than the actual datasets of participants' responses. The top three types of qualitative data that participants are likely to share are (in order): Detailed procedures of data collection (e.g., interview protocols), Survey instruments with actual question items, and Analytic scripts.
- *Perceived technological support* and *extrinsic motivation* are strong predictors for data sharing: the value of these variables can be expected to contribute to a higher frequency of data sharing.
- The variables *intrinsic motivation* and *ease of sharing* are positively correlated with data-sharing behaviors, but were excluded in the final prediction model because they do not significantly contribute to the outcome variance in a regression test.
- Surprisingly, the variables *discipline community practice, data ownership,* and *trust of data quality and that it will be reused* are not found to be associated with data-sharing behaviors.

The findings show that in terms of perceived technology, CS2 participants rated the following higher than CS1 participants: 1) technological supports in data analysis, 2) data collection and 3) preparing data for sharing. However, only the first two are statistically significant according to the Mann-Whitney U Test.

As for the perceived discipline community culture, the U test results again imply that the Case Study 2 participants are more likely to rate higher on "there exists a data-sharing standard" and "people [in the discipline community] care a great deal about data sharing."

When examining the factors that influence data sharing, this study does not find evidence that the three independent variables *trust*, *data ownership*, and *discipline culture* are associated with participants'

data-sharing behaviors via a Pearson correlation. The multiple regression model suggests that variables *extrinsic motivations* and *technological supports* significantly contribute to the outcome variance, whereas *intrinsic motivations* and *ease of sharing* do not.

8.0 CASE STUDY 3: RESEARCH DATA INFRASTRUCTURE IN SOCIAL SCIENCES

8.1 OVERVIEW OF CASE STUDY 3

Case Study 3 (CS3) uses Instrument 3 and reports results based on two focus group sessions and one individual interview with eight employees at the world's largest social science data repository, the Interuniversity Consortium for Political and Social Research (ICPSR). There are two objectives in CS3:

- Objective 1: In order to closely examine data repository services on the support of social science data sharing, it is necessary to gather information about how data professionals carry out current practices at a research data infrastructure. The first objective in CS3 is to capture current practices and functional entities in ICPSR.
- Objective 2: The research questions in this dissertation study are focused on the current challenges of the underlying technological supports and social science data sharing at ICPSR. Therefore, information about current IT practices, barriers to processing social science data, or other challenges are gathered in CS3 to broaden the scope of CS1 and CS2.

Delimitation. In this case study, the Open Archival Information System (hereafter: OAIS) is a scaffolding reference to help visualize current practices and workflow at ICPSR. However, a detailed discussion and evaluation of how ICPSR adopts the OAIS model is eliminated due to this being out of the scope of this dissertation.

8.2 DATA COLLECTION

CS3 comprises two focus groups and one individual interview, all of which were conducted in June 2016 onsite at the ICPSR headquarters in Ann Arbor, Michigan. In total, eight ICPSR employees participated in the study, and seven out of eight were directors or senior managers (at least >10 years). Table 8-1 summarizes the experience (in years) and general responsibilities of the CS3 participants. Group A's session lasted about 75 minutes, Group B's session lasted about 65 minutes, and the individual interview lasted about 40 minutes.

Table 8-1. Participant background

Groups	ID	Year of experience	General responsibilities in ICPSR
A	P01	>10 years	Curation
	P02	>10 years	Curation, data processing
	P03	<10 years	Curation, data processing
B	P04	>10 years	Acquisition, administration
	P05	>10 years	Customer relations, administration
	P06	>20 years	Curation, administration
	P07	>20 years	Administration
*	P08	>20 years	Administration

Note: * Individual interview was conducted.

The topics discussed in these focus groups and the interview were:

Group A — "Curation Services": the emphasis of Group A was on data curation services. Participants include P01 to P03. Figures 1a-1d illustrate a more detailed breakdown of the focus group procedure. In Stage II, each participant wrote on their individual sticky notes and attached them to the whiteboard in the conference room (Figure 8-1a). Individual participants were welcome to write down more notes after interactions or discussions with the other participants in the same

group. Participants were also invited to take advantage of visual aids to elaborate more information about their professional activities (Figure 8-1b). In Stage III, participants added underlying IT and desired IT on the whiteboard using yellow rectangular sticky notes (Figure 8-1c). In Figure 1d, participants continued adding different visual aids, such as the "OpenICPSR" with a dashed line onto the final outcome.

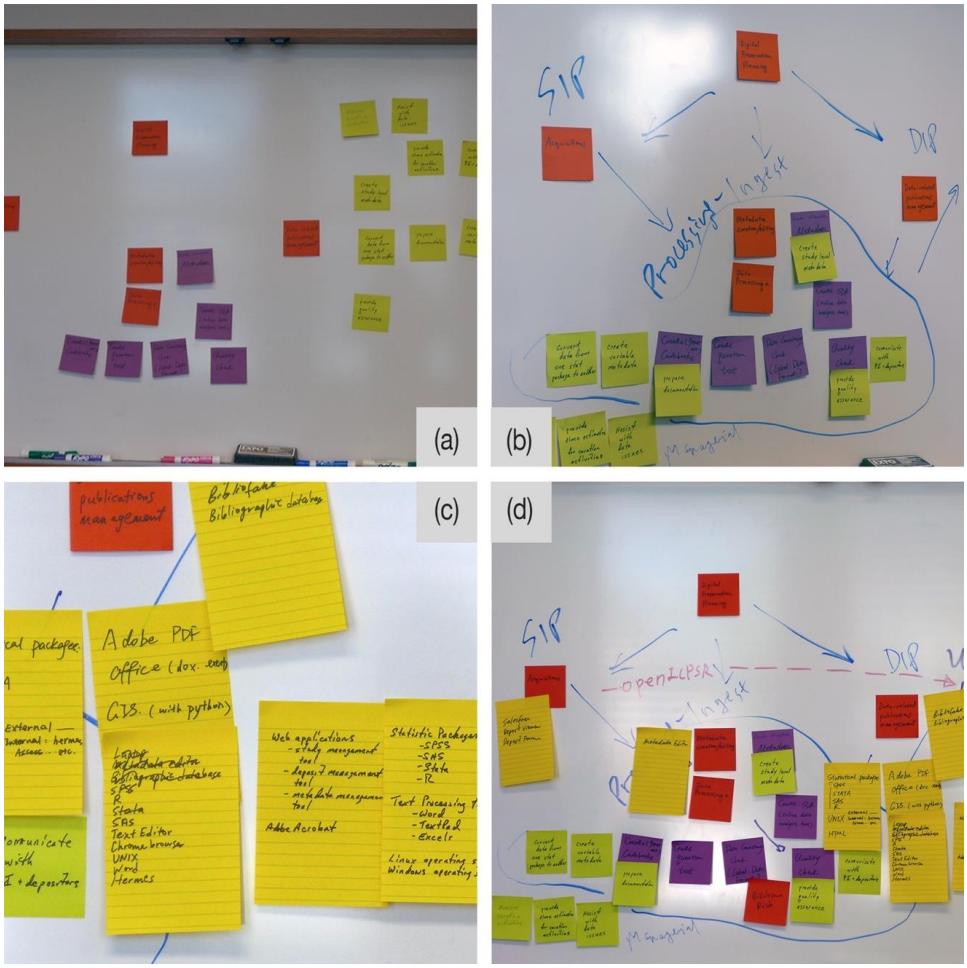


Figure 8-1. Group A activity break-down

Figure 8-2. Group B activity break-down

Group B — "Collection Development": the emphasis of Group B was on collection development and management at ICPSR. All participants in Group B are directors or managers, and their daily responsibilities extend beyond collection development, including acquisition, delivery, supervising, customer relations, outreach, and preservation planning. Participants include P04 to P07 in Table 8-1. A more detailed breakdown can be found in Figure 8-2. First, all Group B participants attached their notes to the whiteboard with no sorting or classification (Figure 8-2a). Afterwards, participants grouped similar activities into columns (Figure 8-2b) and named each cluster themselves

(Figure 8-2c). Note that the focus group mediator did not directly participate in or interfere with participants' sorting process. Finally, as shown in Figure 8-2d, the participants added their IT practice notes onto the white board.

Interview — In addition to the two focus groups, one participant (P08, an experienced director) was interviewed to add valuable perspective and clarify some points regarding the RQs. Questions include:

1. a follow-up on how curation professionals communicate with data depositors about potential disclosure risk;
2. factors that can influence a researcher's willingness to share data with ICPSR;
3. potential challenges and opportunities for social scientists when sharing their qualitative data.

After collecting data from the research sites, all the sticky notes are digitalized and data are entered into a spreadsheet-style table. Specifically, the workflow or cluster created by participants in both focus groups were digitalized by a digital camera. These digital images allowed us to re-create and analyze the focus group results. All conversations that happened during the focus groups and the interview were recorded and transcribed. Participants' quotations on transcription files are managed using ATLAS.ti, a qualitative data analysis software.

8.3 RESULTS

8.3.1 Data curation activities

Since the study collected participants' activities on data curation and collection development at ICPSR, results presented by the participants in Group A resembled the ICPSR Pipeline[3]. However, results presented by the participants in Group B were mostly bottom-up activity clusters, with little similarity to the OAIS structure.

Based on the positions of sticky notes, participant-created activity clusters are integrated with the OAIS model and are presented in Figure 8-3. In Group A's reported activities, after receiving an SIP (submission information package) from the data depositors, data processors perform activities to prepare data for documentation, such as "building metadata" and "creating codebook." The various activities in the data processing stage seem interrelated and not necessarily sequential, as participant P02 expressed, "*once we get everything together, then we start to put all these pieces together and they're all interrelated. You don't have to do one before the other.*"

Unlike Group A's use of a workflow to explain their professional activities, Group B sorts their activities (shown in yellow rectangles in Figure 8-3) into eight clusters: curation, new products, acquisition, outreach, evaluation, management, customer services, and training & education. Group B's clusters overlay with other OAIS functional components except for data processing and metadata building.

[3] ICPSR Pipeline. icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html

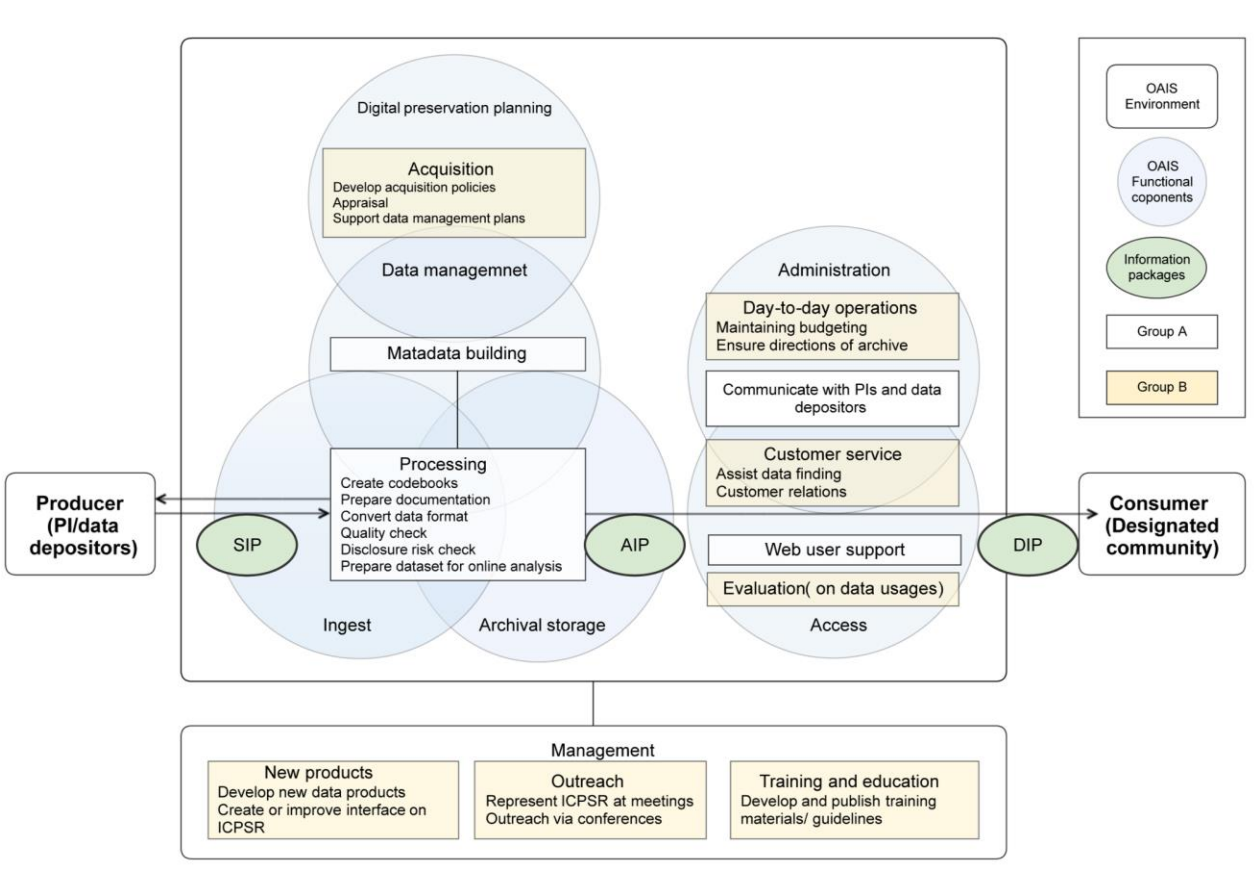


Figure 8-3. Participant-reported activities and OAIS components

As seen in Figure 8-3, only a portion of activity clusters can be covered by a single OAIS function entity. The activities in "Ingest," "Archival Storage" and "Data Management" are overlapping, suggesting that they require support from multiple entities. This is exactly the purpose of viewing OAIS as a reference: although the OAIS model provides a high-level reference guideline, data archives or repositories should expect to work out details and customize the model to reflect their own needs.

8.3.2 Current IT practices

Table 8-2 enumerates the reported technologies based on associated activity clusters. Participants report more IT tools related to "data processing" and their effort to develop "new products". Office software (such as word processors, text editors, and spreadsheets) are the most common tools. On the other hand, participants reported that they prefer Linux-based operating systems in their work environment, and most of their work is done under the Linux environment: "*We do our work in the Linux environment but we have Windows environment that we can also work in as well"* (P02); "(We) log on PC but using Linux" (P01).

Table 8-2. Current information technologies reported by participants

Activity clusters	Current IT	Participants
Acquisitions	Metadata editor, lead management tool, deposit viewer, deposit form, spreadsheet, email	P01, P04, P05
Web team	Bibliographical database (bibliofake), PDF applications	P01, P03
Processing	Word processor, spreadsheet, GIS scripts, SPSS, SAS, Stata, R, text editor, Linux, Windows, Study management tool, deposit viewer, metadata editor, PDF applications, web browser, Unix, Hermes, HTML	P01, P02, P03
New products	Online questionnaire software, usability testing tool, web-hosted service for webinars, responsive design tools, email, Unix, HTML, XML, word processor, funding database, lead management tool, deposit form, email	P04, P05, P06, P07
Outreach	Web-hosted service for conferences, presentation software, Google Analytics, word processor	P04, P05, P07
Evaluation	Text visualization tool, Google Analytics, data mining tools, data visualization tools, online questionnaire software	P04
Management	University financing reporting system, spreadsheet, word processor	P04, P06, P07
Customer service	Email tracking system, web-hosted service for webinars, email, social media, online video	P04, P05, P07
Training and education	Word processor, web-hosted service for webinars, email extension (Boomerang for Gmail)	P04, P05, P06, P07

According to Group A (in which participants used Figure 8-4 to explain the internal workflow of processing an SIP), the process indicates that core activities in the data processing cluster mostly rely on internally developed applications, which include:

- Herme (a file-converting tool that can convert data files from one format to another, such as from SPSS to CSV and SAS),
- Deposit Forms (creating the package after data depositors or PIs finish the deposit;
- Deposit Viewers (allowing curators at ICPSR to view metadata about deposits),
- Metadata Editor, ("creating, revising, and managing descriptive and administrative metadata about a study," [Beecher, 2009, para 5]) A librarian (an employee at ICPSR), "who does all the metadata approval and editing (P01 in CS3)" at this task.
- bibliofake (a database created for storing "bibliographic information and exports it into a format in a system that can use to render that information on the website [P01])."

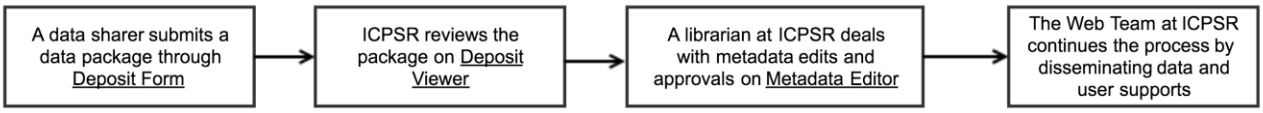


Figure 8-4. The internal workflow of processing data package at ICPSR

Source: P01 hand-drawn during focus group (Group A)

It can be concluded that there is no single integrated platform that handles multiple activity clusters simultaneously. On the other hand, some activities, such as processing, involve more tools and thus are more complex than others. As shown in Figure 8-5, P03 wrote down a couple tools she

used in the process of data package processing. Participant P02 elaborated on what P03 wrote by stating: "*I'm mostly surprised these are all the stuff that we're doing"* (P02).

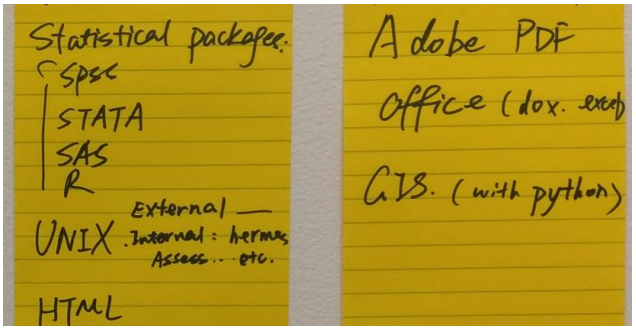


Figure 8-5. A data curator's toolbox for processing data packages at ICPSR (P03)

8.3.3 Desired information technologies

As shown in Table 8-3, Group A precisely describes the tools and technologies needed to address daily challenges. For example, they would like to have technologies that can automatically extract the metadata from an input dataset; as one participant mentioned, "*Wouldn't it be great if there was a form where you uploaded a file and that system would automatically extract all of the metadata for that file" (P01).* They also desire tools that can help "flag" possibly sensitive or harmful content, and technologies that can automatically discover possible identifier combinations. Almost all participants in Group A mentioned the disclosure check: *"You always have to decide, "Is it harmful?" What's the level of harm that's going to happen and what's the level of sensitivity?"* (P02) *"[S]ometimes you miss human sense of what kind of information is dangerous. I know there are tools for disclosure risk but they are not efficient and they cannot identify information [that] we actually identify as disclosure risk"* (P03).

Table 8-3. Current challenges and ideal IT solutions reported by participants

Activity clusters	Current challenges	Ideal IT solutions	Participant
Processing	Metadata are manually extracted.	Technologies that can automatically extract most of the metadata from an input dataset	P01
	Disclosure risks or sensitive content are manually checked	Technologies that can help 'flag' possible sensitive or harmful contents; automatically find possible combination of identifiers	P02 & P03
	Quality control	Tools that can speed up the process for ensuring data quality by checking if file crushes, errors, executing dataset and scripts	P02 & P03
Administration	Hard to estimate "cost" for every single case	Technologies that can estimate needed resources before assigning labor and money.	P06
Management	Hard to synchronize with other departments in the institution	One united and transparent system that can instantly and actively inform or facilitate communication and synchronization between internal departments or separate archives; that can reduce time between contacts	P04 & P06
Training and education	--	A platform that can enhance user engagement and allows customization for training purposes	P05

Since all the participants in Group B are in management positions, their descriptions of ideal technologies are less specific but more comprehensive than those provided by Group A. For example, they desire automated tools to estimate the cost of each study, and systems that can unite multiple departments. Participant P04 called for tools that can "*make things connect and interact across because now we have all of these silos, systems with the University (U of Michigan) with ICPSR.*" She also anticipated this one-stop-shopping system can be developed sooner: *"…the hope is that over the next few years, we'll be putting in a new enterprise system, securities and if this will connect some of those things better or just take one place that you put everything and go in and grab what you need" (P04).*

8.3.4 Barriers and challenges

This section discusses the challenges and opportunities regarding social science data sharing. Table 8-4 lists the challenges and opportunities this study identifies through the focus group sessions (P01-

P07) and the interview with P08. Challenges and opportunities occur at various levels, ranging from individual researchers, their discipline communities and data infrastructures, to the national level. Note that a *cross-level* investigation is needed because a challenge that exists on one level may be solvable by an opportunity existing on another level. Each identified challenge in social science data sharing from data curators' perspectives is explained in the following sections.

Table 8-4. Challenges and opportunities in different levels

	Challenges	Opportunities
Individual level	Social scientists' individual concerns about data sharing: ▪ PI's confidentiality concerns (P01, P08) ▪ PI's confidence of data sharing (P01, P08) ▪ Lack of reward model (e.g., data are not recognized as research products) (P01)	--
Community level	▪ Lack of data sharing standard (e.g., metadata descriptions or file formats) (P01, P02) ▪ Low awareness of data sharing in social sciences (P01, P02)	▪ Data metrics (P01, P03, P05, P06, P07)
Infrastructural level	▪ Labor-intensive process of data curation, especially for qualitative data (P01, P02, P03, P04) ▪ Hard to fulfill various community needs at once (P04, P05, P06)	▪ Active curation (P04) ▪ Enclaves and embargo settings (P01, P02, P08)
National level	▪ Can be both challenges or opportunities: ▪ Regulations and mandates on data sharing at the national level (P07)	

8.3.4.1 *Labor-intensive process of data curation*

Preparing qualitative data for sharing requires extra time and effort. For data curation professionals, open-ended responses can be text-heavy, and the processing cost for time and labor is hard to estimate. For example, participants P01 and P03 had a conversation and described the efforts of processing qualitative responses, "*If you have to read through 10,000 responses"* (P01) – "*Sometimes they*

mention the names, other people name their names or the exact date of something happened, that's the information we don't want them to (reveal)" (P03).

8.3.4.2 *Standard for text data files*

Participants also suggested that it is necessary to adopt and inform data depositors about sustainable digital file formats and standard metadata for qualitative data. Regarding qualitative data curation, ICPSR widely accepts a series of text-based files, whereas the PDF is an exception:

> *"We have a very good handle on that where we put it into an ASCII text file or set ups with qualitative stuff. It's not as cut and dried to use Word as a proprietary format, to use XML, or PDFs, or if you put it in a PDF, is it searchable? (in a rhetorical tone)" (P01).*

8.3.4.3 *Identification of the designated community*

Data curators often face the designated community problem--that is, they find it difficult to clearly identify the target users of a data repository. For example, P06 expressed that they would occasionally ask themselves about who the designated community of ICPSR is: "*there's customers* (research institutions who pay the annual membership fee to ICPSR) *and there's users* (data reusers), *and then people who use our data are often not the people who pay for it"* (P06). Therefore, the team may need to use additional labor and time to repeatedly review potential stakeholders.

8.3.4.4 *Individual concerns around data sharing*

Several observations made by the data curators can help explain why a social scientist might refuse to share data. On the top of the list, social scientists are most worried about "sensitive data" and have "confidentiality concerns": "(One barrier) *is fear of confidentiality or privacy issues, feeling like they have some sensitive information or data that they won't be able to release and so but they don't know about these other*

channels that are available" (P01). In addition, qualitative approaches usually deeply involve the researchers' worldviews; such subjectivity might influence how qualitative researchers view and value their research data, and thus may sometimes result in resistance to archive and share their data. Participant P08, speaking from an administrator's perspective at ICPSR, shared his thoughts on qualitative data sharing and still believes qualitative data sharing is possible: *"data sharing tends to be weakest in qualitative fields because qualitative researchers many of them for various ideological and ontological reasons believe they can't share their data, But it's not true that that's not universal*" (P08).

8.3.4.5 *Community awareness of data sharing*

The majority of faculty and graduate students in social science fields do not share data or are unaware of its importance. Participant P01 related this phenomenon to the low awareness of perceived benefits: "*not everyone or even not the majority maybe know that publishing data or putting your data into a repository is a good thing"* (P01). Other than data sharing, participants also advocated for data reuse from data consumers. P05 stated that she expects to use or develop more publications to raise data consumers' awareness of available data resources: "*I said publications to educate people about-- it's educating for awareness which is different than training how to use data*" (P05).

8.3.4.6 *Reward model for data sharing*

The lack of reward model can be another critical hindrance for researchers' data sharing in general. Participant P01 compared data products with research articles*:*

> *"[Y]ou've probably gone through the tenure process where your reviewers, if you publish a data collection, or let's say you publish an article, but you also spent… a lot of time publishing a data product. That data product is used by thousands of people around the world. That article maybe was*

read by ten people but it was in science or nature, that would be a tenure, the data product, from what I understand, doesn't get nearly the eyeballs or attention" (P01).

8.3.5 Opportunities

Despite the challenges, it can be observed that four encouraging opportunities for social science data sharing from data curators' perspectives. Among these opportunities, data metrics were on the top of the list and were mentioned by participants in both study groups.

8.3.5.1 *Secure dissemination services*

Several participants (P01, P02, and P08) mentioned the enclave policy at the ICPSR. "*We do have a restricted data use policy. People can apply and receive the data from our secure downloads if they can have it or if it's just really restricted, we can put it in a physical enclave or we have a digital enclave where people can log into it and only use the data there.* (P02)" Research data infrastructure also pays attention to the potential disclosure risks, and data repositories such as ICPSR often offers secure dissemination services. Such security mechanisms are an opportunity to address the individuals' confidentiality concerns mentioned above.

8.3.5.2 *The scholarly recognition and the maturity of data metrics*

Despite imperfections, citation-based bibliometric methods have been widely used to evaluate scholars for promotion, tenure, hiring, or other recognizing mechanisms (Borgman, 2007). However, data citation or data publication is not a common recognizing mechanism in academia.

After being asked why social scientists would share their data to ICPSR, P01 stated,

"I heard someone talking about data citations or will it be an encouragement if your data got cited. It gives you credit as your paper is cited. I think that will be a good idea or encouraging for people" (P01).

In CS3, participants in both focus groups repeatedly mentioned the lack of recognition of data citations:

"It's funny that you look at the citation or reference of a book or a journal article and that's very well established in research and academia but this you can't say nearly the same for our data collection. It's not yet considered a first-rate research product and as a result it affects other aspects of the research life cycle" (P01).

Although NSF (2013) has recognized data as a research product since 2013, it is still taking time for academia to form an agreement and adopt data publications as research products (Costas, Meijer, Zahedi, & Wouters, 2013). To encourage data sharing in social sciences, the community can consider data sharing a kind of academic contribution by adopting data metrics. P05 in Group B expressed her positive attitude about the connection between providing data metric services and a PI's willingness to share data at ICPSR:

"… individual PI, they might be excited to see downloads and citations and search…They can say, look at how much impact we have had… [B]ut again it's all still relatively new" (P05).

8.3.5.3 *Call for an "active curation"*

To speed up the process of data curation, participant P04 mentioned the concept of *active curation*, a new model of accomplishing data curation piece by piece (Myers et al., 2015). The traditional curation model usually requires everything to be available before proceeding to the next step, whereas active curation is an incremental model where metadata and elements can be added over

time: "*That's where my wishes came from, reducing the time it takes to get data in the door, supporting active curation, so maybe we can get the data in before they have to actually deposit it or let others use it, but if we can help them along the way"* (P04). This opportunity not only reduces curation time, but also ultimately allows PIs to proactively update their datasets. This is beneficial for PIs who are hesitant to share data because they are afraid that errors or mistakes in their data will be pointed out.

8.3.5.4 *Call for a national policy*

Participant P07 mentioned the UK, which has national policies that encourage UK researchers to submit datasets to the national archives: "*Yeah, and many other countries like UK, there is requirement that people deposit their data in a particular place.* (P07)." There is no national-wide data sharing infrastructure as of 2016 in the U.S., and there is no universal guideline for selecting a data archiving platform. The existence of a national policy can simplify PIs' effort to select a data archiving platform, but it would be challenging to build the supporting infrastructure for such a policy.

8.4 SUMMARY OF CASE STUDY 3

Through two focus group sessions and one individual interview with eight total ICPSR employees, CS3 examines data professionals' current practices and IT practices at ICPSR, a leading social science data repository.

In summary, CS3 showed that 1) the cost of preventing disclosure risks and 2) lack of agreement on a standard text data file are the most apparent obstacles for data curation professionals who handle qualitative data; 3) the maturity of data metrics seems to be a promising solution to several challenges in social science data sharing.

Based on participants' points of view, several challenges and opportunities for data sharing in social sciences are observed. The reported findings reveal several challenges in social-science data sharing, such as data ownership and confidentiality concerns; although, again, a particular challenge may exist on one level (e.g., PIs' concerns about data sharing at the *individual* level), but would be resolvable by an opportunity existing on another level (e.g., the maturity of data metrics at the *community* level). Data sharing and curation in social sciences remain challenging to scale due to privacy concerns and a labor-intensive process, especially with regard to qualitative data sharing. Better and automated tools would be required to help detect or perform disclosure check.

One future work that can be extended from CS3 is to compare its results with related work based on the investigation on social scientists' data sharing and reuse practices (e.g., Yoon, 2016; Curty, 2016). A cross-level (i.e., individual, institution, community, and infrastructure) triangulation is exceptionally needed for capturing the whole picture of data sharing and reuse practices in social science. Another future direction is to compile a list of design principles to improve the design of a data curation system, based on the collected IT practices and ideal technologies in this study.

9.0 DISCUSSION

This chapter discusses the findings of all the studies in this dissertation— two preliminary studies and three case studies—and triangulates the connections among them. Following the research framework proposed in this dissertation, this chapter highlights eleven discussion points, as summarized in Table 9-1.

Table 9-1. Roadmap of discussion points and related framework

Index	Result discussion points	Dimensions to studying data-sharing practices	Framework to support digital scholarship	
			Knowledge Infrastructure (KI)	Theory of Remote Scientific Collaboration (TORSC)
Ch 9.1	▪ Data sharing in discipline repositories ▪ Research activities and data sharing	Data sharing practices	--	--
Ch 9.2	▪ Confusion about data ownership and its research value ▪ Sharable qualitative data	Data characteristics	▪ Artifacts	▪ The nature of the work
Ch 9.3	▪ Discipline community practices ▪ The funder policy ▪ The call for establishing best practices	Organizational context (specializing in discipline community)	▪ Institutions (organizations) ▪ Routines and practices ▪ Policies	▪ Common ground ▪ Management, planning, and decision making
Ch 9.4	▪ Perceived benefits for data sharing ▪ Norms and concerns: confidentiality in qualitative data	Individual motivations and concerns	▪ People (individuals) ▪ Shared norms and value	▪ Collaboration readiness
Ch 9.5	▪ Technological readiness toward the data sharing culture ▪ Ideal technologies for data sharing-reuse cycle	Technological readiness	▪ Built technologies (system and networks)	▪ Technology readiness

9.1 THE LANDSCAPE OF DATA SHARING IN SOCIAL SCIENCES

9.1.1 Data sharing in discipline repositories

This dissertation study confirms that data sharing is still limited in social sciences. The triangulated result indicates that the majority of social-science faculty members and students do not share data or are unaware of its importance (in Table 9-2). Early-career social scientists in PS1 and CS1 seldom share their data along official channels, such as institution repositories or discipline repositories, even though they highly value data sharing and witness data sharing in their fields.

Table 9-2. Triangulations on low awareness of data sharing

	Justifications		
Main message	Preliminary study 1	Case Study 1	Case Study 3
The majority of faculty and students in social science fields do not share data or are unaware of it.	All participants indicated that they are willing to share upon request. Few of them have experiences of sharing data in data depositories.	The insufficient activities in both manuscript sharing and data sharing.	"still not everyone or even not the majority maybe know that publishing data or putting your data into a repository is a good thing. (P01, CS3)" "it (data sharing) seems like this a big thing and it's getting bigger around the world, but then we talk to majority of students and professors and other people who aren't in this field act "Oh, what is that? Oh really?" It's... I don't know, it's strange. (P02, CS3)"

The results of CS3 reveal the low awareness about data sharing in social science students and faculty. The participants in CS3 attribute this low awareness to the lack of reward models (i.e., inadequate awareness of perceived benefits). Other possible explanations of low awareness include the fact that data-sharing mandates did not exist until the 2010s. Moreover, social scientists rarely receive formal training in data curation and management, not to mention data sharing. Jahnke et al.

(2012) observed that out of the researchers they studied, none "had received formal training in data management practices."

For those who are aware, such as the participants in CS2, there is clearly a lack of best practices and awareness of standards regarding data sharing in social sciences. Further details are discussed in Section 9.3.

9.1.2 Research activities and data sharing

Both Preliminary Study 1 (PS1) and Preliminary Study 2 (PS2) demonstrate different patterns of participants' research processes and methods, which motivated the design of related questions in Case Study 1 (CS1). However, based on the CS1 responses about data-related research activities and participants' data-sharing behaviors, no statistical difference was found between qualitative, mixed, and quantitative methods. For example, an ANOVA test on the results of CS1 suggests that researchers whose preferred method is quantitative data report more frequent publishing activities than the other two methods, whereas other data production activities are not significantly different. In CS2, there is also no difference between the data-sharing behaviors of qualitative and quantitative researchers. Although social scientist participants in this dissertation study responded differently in the way they conduct their research in PS1 and PS2, there is not a statistical difference between research methods when it comes to decisions about sharing data and actual data-sharing behaviors.

Another similar observation is that there is no significant difference among disciplines throughout all the studies in this dissertation. Although disciplinary difference is observed in researcher data production in both PS2 and CS1, there is no evidence to conclude that disciplines are a factor affecting data-sharing behaviors.

In summary, one repeated finding is that although qualitative and quantitative researchers are different in many aspects based on the preliminary studies, they resemble each other when it comes to manuscript sharing and data-sharing frequency (in CS1 and CS2). One possible explanation for this is that there are shared internal and external drivers (or barriers) faced by most social scientists. Such shared factors include data ownership, funder pressure, and ethical considerations.

9.2 DATA CHARACTERISTICS: THE NATURE OF THE WORK

In the context of data sharing, the nature of qualitative data can be mapped to "the work" and "the artifact" in the theories of KO and TORSC. This section highlights the discussion of two issues related to research data that social scientists interact with: 1) social scientists' confusion about data ownership and its value, and 2) the gap between the sharable data perceived by social science researchers and the shared data expected by policy makers.

9.2.1 Is that "my" data? Confusion about data ownership and its research value

Table 9-3. Triangulation on data ownership and research ownership

Main message	Justifications	
	Case Study 1	Case Study 2
Participants are concerned about the confusion and uncertainty of data ownership and its research value.	"[I]f I download government data, but select a subsample, clean up the coding, and create some new variables, is that "my" data? In my field, we would consider that to be your own, but there's not huge value in sharing that when the primary source is publically available unless someone is trying to replicate your results. (P24)"	The hypothesis of "perceived data ownership would positively influence data sharing behaviors" is not supported.

This dissertation study found that data ownership and perceived research originality are critical points to be considered and clarified before researchers share data. The results of this dissertation echo the findings in several prior studies, which alludes to the complexity of data ownership and research originality. Such complexity can be viewed from three aspects.

First, data ownership is a major concern raised repeatedly by participants in the open-ended responses in CS1 and CS2. The fact that many participants are confused and uncertain about data ownership shows that social scientists may hesitate to share data without knowing which party possesses and has responsibility for it. The triangulated result is listed in the matrix table in Table 9-3.

Second, a participant in CS1 mentioned that he is not sure if his research data has original value— "*is that my data?"*—because what he did was "*download government data, but select a subsample, clean up the coding, and create some new variables*" (P24 in CS1). This finding is consistent with related work. For example, Jahnke et al. (2012) note that some participants in their interviews "wondered who might be interested in their data" (p.11). Curty (2016) also remarks that some social scientists believe that their research outcome might be overlooked or undervalued.

Third, for the CS2 participants with prior experience sharing data, the hypothesis of *perceived data ownership positively influencing data-sharing behaviors* is not supported. One possible explanation for this is that data ownership is more likely to be a threshold condition than a correlation: a PI must clear the claim of ownership before one is able to share data; however, the sharing behavior does not depend on what the perceived data ownership score is.

In summary, data ownership is challenging because 1) it is unclear whether the data belongs to the researchers, the informants, or funding agencies; 2) the level of originality (i.e., whether the data is qualified to be called "their own data") is also questioned when the data is collected from

third-party resources. Addressing both issues should be the top priority when developing the best practices for data sharing.

9.2.2 An oxymoron: sharable qualitative "data" is not data

Most participants agree that the majority of shareable qualitative data are instruments or research tools such as protocol. There is not yet a consensus about sharing actual empirical data. In actuality, these methodology-related documents or tools, broadly speaking, are part of research data. However, they are *not* data when considering the strict definition provided by the U.S. federal government[4], in which the data should be "necessary to validate research findings." Although further study is needed to unveil why qualitative researchers prefer sharing research tools over actual data, several conjectures can be made here.

One possible explanation recalls the philosophical considerations of qualitative studies: qualitative approaches usually deeply involve the researchers' subjectivity, which shapes how they value and explain outcomes. Therefore, as some researchers have noted, "qualitative data are researcher-centric, gathered in connection with a specific inquiry, and used just once" (Elman, Kapiszewski, & Vinuela, 2010, p.24); sharing research instruments and protocols is more compatible, as different researchers can use such instruments to gather their "researcher-centric" data. Similarly, participants in CS3 observed that qualitative scholars rarely share data:

[4] The definition of "research data" is "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (OMB Circular 110).

"data sharing tends to be weakest in qualitative fields because qualitative researchers many of them for various ideological and ontological reasons believe they can't share their data" (P08, CS3).

While there are rich studies on the topic of withholding data in STEM fields (e.g., Compell, 2002; Krawczyk & Reuben, 2012), other possible explanations may be applied to the context of social sciences and qualitative data, including 1) higher expected reward or impact for sharing tools rather than data, because tools can be applied to a wider range of research and scenarios; 2) worries about informants' confidential information being revealed; 3) fear of the research validity and reliability in their qualitative or mixed method studies being criticized. However, further study is needed to verify these possible reasons.

In addition to the benefits of data reuse and teaching, sharing raw data can also encourage a rigorous research process because researchers need to demonstrate how they undertake data production. Therefore, one downside of only sharing research tools (and withholding actual empirical qualitative data) is the decrease of research transparency. To overcome this, researchers whose actual empirical data is unshareable should still consider sharing templates or examples of actual empirical cases.

9.3 ORGANIZATIONAL CONTEXT

Both KO and TORSC theories consider contextual aspects around researchers and their work environment. Such environments may be institutions, policies, organizational routines, and operations. Particularly, both theories mention the concept of "common ground" (Olson, J. & Olson, G., 2013; Edwards et al., 2013, p.6), which represents a shared context, such as shared knowledge and shared

practices. However, unfortunately, this dissertation does not find enough evidence of the influence of common ground in qualitative data sharing.

This section discusses the lack of common ground from three aspects: 1) it is unclear how much influence the discipline community has regarding social science data sharing; 2) the funder's policies or attitudes are crucial in determining whether social scientists share data; 3) the participants in this dissertation study stressed the need for best practices.

9.3.1 Discipline community practices

The results in CS2 suggest that the discipline community's data-sharing practices do not play an important role in data sharing. In other words, based on the perception of the participants in CS2, a social scientist does not have higher data-sharing frequency if one perceives that the community has better data-sharing practices.

One possible explanation for this is that most of the CS2 participants are senior professors or researchers who are more independent, therefore their behaviors are less likely to be influenced by the community. Although this dissertation finds no evidence of dependency between discipline community practices and individual social scientists' practices, further study is required to clarify the role of a community in data sharing. Section 10.2 further discusses possible roles played by the community.

9.3.2 The funder's policy

This dissertation study finds that policies about data sharing from the funders of a project can influence researchers' data-sharing behaviors. The policy can be both strongly or fairly positive (e.g.,

mandates or encouragement) and negative (e.g., imposing restrictions). In both CS1 and CS2, several participants mentioned that funder policies play an important role in data sharing (Table 9-4).

Table 9-4. Triangulations on funder's policy

Main message	Justifications	
	Case Study 1	Case Study 2
The funder might be the one deciding whether to share research data, reducing the level of research autonomy.	"my funded research is in the field of evaluation where much of our work is sponsored by clients so it is very challenging share data. (P66 in CS1)"	I work on a NIA-funded study…I HAVE to share my data and it doesn't matt[e]r if I have enough time, money, etc. to do so. (P128 in CS2)" "Data sharing in many instances faces significant challenges where the research is funded by private entities or institutions that seek to use such outcomes for own programming. On the flip side, a number of research initiatives funded largely for public go[o]d/use often have less restrictive environments for sharing. (P132 in CS2)"

One participant described a mutually dependent relationship with the funders:

> *"Data sharing in many instances faces significant challenges where the research is funded by private entities or institutions that seek to use such outcomes for own programming. On the flip side, a number of research initiatives funded largely for public go[o]d/use often have less restrictive environments for sharing." (P132, CS2)*

That is, the funder might be the one deciding whether to share research data, reducing the level of research autonomy.

While the participants in CS1 and CS2 indicate the importance of funders, prior work has found no causality between funder pressure and data-sharing behaviors. Specifically, Kim and Stanton (2012) hypothesized that the pressure from funding agencies and journal publishers would influence social scientists' data sharing. However, they find no statistical evidence supporting this

hypothesis. Since this dissertation and Kim & Stanton's study have different research samples and directions, further work is needed to examine the root cause of these inconsistent interpretations.

9.3.3 The call for establishing best practices

The call for establishing best practices or standards has gained considerable momentum. Multiple researchers stress that it is time to establish best practices as well as a standard for sharing data in social sciences. This inadequacy is universal irrespective of which research methods they preferred: it was observed in researchers who preferred qualitative, quantitative, and mixed method data sharing (Table 9-5).

Table 9-5. Triangulations on the call for best practices

Main message	Justifications	
	Case Study 1	Case Study 2
The call for establishing data sharing best practices/standard in social science has gained considerable momentum	"I think I'd be happy to share data and code more frequently if I had a better sense of good practices. (P46, CS1)"	"It's time to establish best practices & resources to support data sharing. (P114 in CS2) "

One participant in CS1 said, "*I think I'd be happy to share data and code more frequently if I had a better sense of good practices*" (P46, CS1), whereas another participant stated, "*It's time to establish best practices & resources to support data sharing*" (P114, CS2).

9.4 INDIVIDUALS' READINESS: MOTIVATIONS, NORMS, AND CONCERNS

This section focuses on social scientists themselves—the concept of the "individual" in KO and TORSC. "Individuals" or "people" become an essential aspect when examining data sharing in social sciences. While a successful environment should provide incentives and help eliminate barriers for individuals to share data, induvial readiness and motivations are also crucial factors. This section discusses the perceived benefits and barriers that might encourage or discourage the data-sharing behaviors of individuals.

9.4.1 Perceived benefits for social scientists

Table 9-6. Comparisons on perceived benefits across case studies

Inquiry	Comparison		
	Case Study 1	Case Study 2	Case Study 3
Motivations to share data	Seeking collaboration opportunities and helping others	Making an impact on research for and teaching next generation (citation increase, impart the social research method) and helping others (fulfill others' research needs)	Citation increase

While comparing different parties' motivations to share data (Table 9-6), this dissertation study found that participants in CS1 and CS2 overwhelmingly have the highest averages in intrinsic motivation. Interestingly, for extrinsic motivation, participants in CS1 identified "seeking collaboration opportunities" as the main one, whereas CS2 participants have significantly higher ratings for "gaining more citations" than participants in CS1. This observation implies that while

CS2 social scientists with data-sharing experience care about altruism, they also care about traditional scholarly recognitions such as gaining citations, more than the CS1 junior social scientists do.

The above observation in CS2 has been triangulated in CS3. Data curation professionals were asked about their practical observations on the factors influencing social scientists' willingness to share data, and they perceived increased citations as a benefit.

As mentioned in the results of CS3 (Section 8.3.5), the citation-based bibliometric in journals has been widely adopted to assess researchers for hiring, tenure, promotion, or other recognition (Borgman, 2007). Consistent with Costas et al. (2013), there is a need to reconsider the reward system: if sharing data can effectively return as rewards (e.g., increased credits or rewards in the reviewing or promotion processes from their institutions), it may take shorter time for the academics to embrace a data sharing culture.

9.4.2 Norms and concerns: confidentiality in qualitative data

While this study confirms that technology and extrinsic motivations are drivers for sharing qualitative data, confidentiality concerns and labor-intensive processes remain major barriers, as observed in related work (Chapter 2) and confirmed in CS3.

Since social science studies often rely on close relationships with participants, confidentiality concerns might outweigh the benefits of data sharing. This dissertation study repeatedly discovers that social scientists are worried about "sensitive data" and have "confidentiality concerns" about sharing data. This can be triangulated across CS2 and CS3. Table 9-7 below provides evidence for this triangulation. These observations are consistent with related work, in which researchers discuss PIs' challenges through the process of sharing qualitative data:

Table 9-7. Triangulation on confidentiality concerns and efforts

Main message	Justifications	
	Case Study 2	Case Study 3
Since social science studies often have close relationships with the participants, confidentiality concerns might outweigh the benefit of data sharing.	"Confidentiality and deductive disclosure are huge issues for me [a]re: data sharing, since all of my research is about risk behaviors (sexual assault, dating violence, sexual activity, substance use) and much of it involves minors... (P86 in CS2)" "I have only deposited data because it was required by federal grants, and even then was hesitant due to confidentiality concerns. (P73 in CS2)	"(One barrier) is fear of confidentiality or privacy issues, feeling like they have some sensitive information or data that they won't be able to release and so but they don't know about these other channels that are available. (P01 in CS3)"
Time and labor are invested for ensuring good quality of data description and metadata.	*"It took us one year to prepare data to upload to ICPSR - it was not simply ensuring good descriptions or accurate metadata but just ensuring that the files were complete, non-redundant and interpretable."* (P106 in CS2)	"For qualitative data, what we have to do is sometimes we have to read through all the responses (for a disclosure risk check)" (P03 in CS3)

> *"A researcher wanting to safely observe both sets of considerations, whose only guidance on the issue might come from a local, risk-averse, and tradition-bound institutional review board, will almost always conclude that sharing of the granular data they have collected in interactions with human participants is not a good idea and will thus perpetuate the status quo of putting all these rich materials "under lock" or, even worse, promising to destroy them at the end of the project."* (Bishop, 2009, p. 261, as cited in Karcher, Kirilova, and Weber, 2016*).*

In order to protect participants' privacy and sensitive information, researchers need to perform additional operations (e.g., informed consent, deducting real information, anonymization, converting specific information to general information, performing disclosure checks, etc.) throughout the process of data production, sharing, and reuse. These operations are labor intensive, as pointed out by one PI in CS2: "*it was not simply ensuring good descriptions or accurate metadata but just*

ensuring that the files were complete, non-redundant and interpretable" (P106). Since sharing qualitative data consumes extra resources and time, it is more challenging to share than quantitative data.

9.5 TECHNOLOGICAL READINESS AND INFRASTRUCTURE

This section discusses the technological readiness perceived by social scientists and their expectation of ideal technologies. This is closely related to the concept of "technologies" in KO and TORSC.

9.5.1 Technological readiness toward a data sharing culture

Guided by CCMF, this dissertation unveils that the social science community exhibits slow adoption of certain technological mechanisms, including data identifiers (mentioned by all disciplines in PS1), data metrics and impacts (mentioned by anthropologists and political scientists), as Figure 4-4 shows.

Moreover, CS1 and CS2 suggest that social scientists lack awareness about technical standards such as DDI. As shown in Table 9-8, evidence can be found in the statistical data in CS1 and CS2: only 14% of CS1 participants agree that there is a standard for data sharing in social sciences, but even CS2 participants who had shared data before yield only a 31.1% agreement. There is no doubt that both researchers and information professionals should pay closer attention to developing best practices or advocating for data sharing in social sciences. Unfortunately, despite the maturity of DDI, most of participants were unware of the standard. To address this issue, the community is obligated to advocate such standards and educate early-career researchers.

Table 9-8. Triangulation on technological readiness on standards

Main message	Justifications		
	Preliminary Study 1	Case Study 1	Case Study 2
Technical standards (data description, identifier, metrics) are the weakest link in social sciences	Several items related to technical standards (e.g., 5.6 data identifiers, 2.12 data metrics and impact) are rated least developed in PS1.	Only 14% of participants agree that there is a standard for data sharing in social sciences	31.1% of participants agree that there is a standard for data sharing in social sciences

The community can help improve technological readiness on technical standards in two aspects: advocacy and training. To support the development of best practices, it is necessary to establish systematic training ranging from data production, curation, to sharing. Such training should improve early-career social scientists' awareness about data sharing and reuse.

9.5.2 Ideal technologies for data sharing-reuse cycle

The challenges regarding underlying technology include 1) uneven technological support throughout the data lifecycle, 2) lack of coherent practices, and 3) slow technological evolution to support management of research products.

First, in CS1 and 2, social scientists rate technology or resources unevenly throughout the data lifecycle: tools for data production tend to be considered sufficient, while tools related to data sharing and reuse are rated insufficient. There are two possible interpretations of this. On the one side, it reflects there is truly a lack of research data sharing and reuse technological support. On the other side, these uneven scores may stem from the low awareness about data sharing. If researchers are not informed or aware of data sharing when they conduct research, it is very natural for them to overlook its existing support. For example, in Preliminary Study 2, most participants' visualizations (five out of eight) do not cover activities related to data sharing or even publishing.

Second, current IT practices are customized and their coherence needs to be improved. The findings of CS3 reveal that data processing activities in ICPSR have been handled by internally developed tools, which is consistent with the observation in PS1. That is, social science projects tend to require a unique set of IT functionalities, and thus it is common to develop customized tools for a specific task rather than using general-purpose tools for multiple tasks.

Consequently, since tools are scattered, researchers may need to exert extra effort to adapt themselves to the workflow by using separate tools. In CS3, data curation professionals expect a more harmonized platform on which people can work together smoothly. However, not every participant in CS3 elaborated on the desired IT's possible functionalities and appearance, so a future specific participatory study is anticipated to capture more details.

In sum, the current technological supports in social scientists' work environments are either lacking specific functions in certain research stages or lacking a coherent set of structures and management. Therefore, ideal technologies should seamlessly support a social scientist throughout the research data lifecycle: a better tool on which social scientists can manage most qualitative data, artifacts, records, instrument protocols, and research products generated by the blooming and diverse research methods in social sciences. Balancing the functionalities between "allowing-diversity" and "being coherent" in designing such a technique is key to advancing qualitative data sharing practices.

10.0 IMPLICATIONS AND CONCLUSION

This chapter considers the implications, including theoretical implications and managerial implications, of this dissertation study.

10.1 THEORETICAL IMPLICATIONS

This dissertation study developed a research framework by incorporating Knowledge Infrastructure (KI) and the Theory of Remote Scientific Collaboration (TORSC). The result findings have several implications for this study's design of research framework, as well as KI and TORSC.

10.1.1 An interwoven scholarly infrastructure

10.1.1.1 *The work environment*

In the designs of Instrument 1 and Instrument 2, the institution, department, and discipline communities are often interwoven in the research context; thus, it is hard to precisely categorize questions regarding technological infrastructure, organizational culture, and research culture. Although the theories of KI and TORSC can be applied to individual organizations, they fall short when encountering interwoven disciplines and institutions—that is, participants from a variety of sub-disciplines (in CS1 and CS2) or from different organizations (in CS2). For example, particular

supports like funding resources or technological resources can be obtained by researchers either from external funders (e.g., from a discipline community or a national funding agency) or from the local institution or department.

10.1.1.2 *Technology and human resources*

Sometimes it is hard to clearly separate technology from human resources or human-made static resources (such as Libguides), because most of the time people may be required to work together with technology. For example, a librarian holding a workshop on data cleaning tools can be viewed as either a technological support, a human support, or an organizational support. Practically speaking, a precise categorization of the above-mentioned support is very difficult to achieve, based on the research practices in this dissertation study.

10.1.1.3 *The strengths and limitations of TORSC and KI*

This dissertation study leverages the strengths of TORSC and KI while identifying and working around their limitations. TORSC and KI are powerful theoretical frameworks for data sharing research because they 1) systematically review data-sharing practices, covering most of the attributes, and 2) can flexibly create multiple instruments, such as profile tools, questionnaires and focus groups.

However, while TORSC and KI can roughly describe the discipline community, technological infrastructures, and the ecosystem of an organization by ethnographically profiling researchers' sharing behaviors, one critical limitation of TORSC and KI is that they are unsuitable for categorizing research context when applied to self-mediated questionnaires or self-mediated profile tools. Therefore, in addition to using survey methods (profiling, questionnaire, interview, or

focus groups), future work can strengthen the study results by introducing ethnographic observation approaches to fully utilize the advantages of KI and TORSC.

10.1.2 Implications for data profiling tools

Some questions in Instrument 1 (CS1), borrowed from the profiling tools (e.g., CCMF and DCP), are context-specific. For example, data volume (the totality size of data in a project), data sensibility, and data shareability can vary significantly depending on the projects themselves. Another example is the research stage of a project. In a real-world situation, a researcher might work on multiple research projects in parallel: some projects might be closed, whereas others might still be in early stages and not ready for any form of sharing. Since the situations can differ from project to project, it is imperative to ask the participant to focus on one completed project when reporting on a cross-sectional study. Specifically, for Instrument 2, participants were asked to recall one of their most representative projects when they answered the questions. However, this approach might risk losing generalizability, because it limits the survey results to one single research project. Striking the right balance between providing context-specific questions and preserving the generalizability of a survey is difficult to achieve.

Another example of losing specific context is found in CS1 and CS2, where participants were not asked to identify any ownership conflict claims (e.g., conflicts between institution vs. researchers, informants vs. researchers, or sponsors vs. researchers). Although it might be helpful to know the types of ownership problems, in practice, it is difficult to collect information in such granularity in a self-mediated profile tool or questionnaire.

10.2 MANAGERIAL IMPLICATIONS

This section highlights several managerial implications that offer actionable remarks and suggestions for further data sharing research and practical service sectors. The managerial suggestions are summarized below in Box 1.

10.2.1 Researchers who handle qualitative data

The main points derived from this dissertation repeatedly reveal the sensitivity, complexity, and heterogeneity of qualitative data. Although it might be too early to conclude the best practices of qualitative data sharing, the findings show that experienced data sharers think it is more likely to be possible to share methodology-related instruments than the raw data that leads to the research results. Besides teaching researchers how to best anonymize data, it might be beneficial to also help them identify sharable data (e.g., protocols, instruments, or research tools) during data production and how to claim data ownership[5].

[5] There are several federal resources about the discussion of data ownership claims. For example, as cited in the U.S. Department of Health and Human Service Office of Research Integrity (n.d.), Loshin (2002) clearly identified a range of "*possible paradigms used to claim data ownership*". These claims of data ownership are based on different degrees of involvement in or contribution to the research endeavor. Such claims include several parties such as the creator (who generates data), organization, or funder ("*the user that commissions the data creation claims ownership*").

Box 1. Managerial suggestions to different stakeholders

For researchers who handle qualitative data:

- Explicitly inform participants about data sharing. If possible, the researchers should inform the participants of potential data sharing in the consent form. If participants are unable to sign consent forms, the researchers should carefully evaluate the risk of sharing data.
- Remove any identifiable information in the shared data. Researchers should anonymize and de-identify the shared data to protect the participants' identities and privacy.
- Provide an example when raw data is unshareable.

For institutions:

- Strengthen technological supports for data sharing.
- Incentivize data sharing. To do this, institutions can consider data metrics and citations as an additional indicator for promotion, since data sharing not only helps advance research but also serves the community.
- Immerse early-career social scientists in the data sharing culture. To cultivate data sharing, institutions can engage and expose early-career social scientists (i.e., senior graduate students, post-doctoral researchers, and assistant professors) to trainings on data transparency and the spirit of open research.

For discipline communities, journal publishers, and data infrastructures:

- Provide guides and best practices. Discipline communities and data infrastructures can investigate discipline-level best practices, and professional associations can also provide data sharing guides. Such guides can help researchers prepare data sharing anonymization and select data types.
- Incentivize data sharing (at the institutional level).
- Advocate discipline repositories and existing metadata standards.
- Encourage the sharing of tools, coding results, and selective transcripts. Journal editors should acknowledge alternatives to sharing raw data, allowing tools, coding results or selective interview transcripts to be shared as an alternative to a full set of interview transcripts.

For national policy makers:

- Formulate flexible policies for qualitative data sharing. One policy cannot fit all. Policy makers should consider a "minimal standard" for sharing qualitative data, as sharing research tools or selective records is better than sharing nothing.
- Investigate the balance between privacy and transparency.

Therefore, for researchers who share qualitative data, one best practice concluded by this dissertation is to protect informants while simultaneously ensuring research transparency. Note that data ownership must be cleared and claimed before any form of sharing. Researchers should know how and when data will be shared and include those statements in the consent form.

The following are two strategies proposed by this dissertation study.

- Full disclosure: if a researcher decides to share actual data (e.g., interview transcripts and questionnaire responses) of the participants, one should carefully anonymize personal information and identifiers linked to informants to prevent any and all disclosure risks. Many de-identification techniques regarding anonymization of qualitative material are in practice, including using a pseudonym, reducing the precision of information, removing direct identifying details, generalizing the meaning of detailed text, and using a vaguer descriptor (QDR, 2012; UK Data Archive, n.d.). A researcher needs to replace all the identifiers within the research data. Most importantly, qualitative scholars should document and keep the anonymization records carefully. Table 10-1 provides an example anonymization log for qualitative data de-identification.

Table 10-1. Example anonymization logs for anonymizing qualitative data

File index	Page index	Original (real information)	Change to	Justifications
Transcript #1	p.1	Leah	Emily	Using a pseudonym for the real name
	p.2	Age 29	Late 20s or age range 20-30	Reducing the precision of information
	p.2	Interviewed on March 27	Interviewed in March	
	p.4	Pittsburgh	City in the East Coast of the U.S.	Removing direct identifying details
	p.4	Main branch, Carnegie Library of Pittsburgh	Main branch of the city library	
	p.5	Director, Digital Strategy & Technology Integration	Leader in technology-related services	Generalizing the meaning of detailed text
	p.8	Amy	My colleague	Using a vaguer descriptor

Source: The anonymization protocol is recreated based on QDR, 2102 and UK Data Archive (n.d.)

- Partial disclosure: In some cases, research data might not be able to be completely anonymized, "anonymization would lead to too much loss of content or data distortion" (QDR, 2012, p.6), or hard to use for a potential secondary analysis. Setting an access restriction such as an enclave policy at ICPSR can be considered in such cases (See Section 8.3.5.1 Secure dissemination services). If the actual empirical data is not totally suitable for sharing, or the anonymization process would place an unreasonable burden on a researcher, the researcher may only share research instruments and the coding or analysis results. A few examples of real responses can be provided and appended to the research instruments. Through such examples, data reusers will know how to better reproduce the study or validate analysis results.
- Data regarding potentially vulnerable individuals: sometimes a social scientist may deal with data involving potentially vulnerable individuals such as minors, patients, people with special economic status, prisoners, and so on. One should approach these participants from the same standpoint as they would adults or the general public, but be particularly

careful to explain the risks involved and "potentially morally harmful effects" for participants (as suggested by Morrow, Boddy, & Lamb, 2014, p.11), and be sensitive to ensure all possible combinations of traits that could identify the sensitive group are eliminated.

These suggestions are also applicable for the current universal data management and sharing policy. This dissertation study also suggests that funders or institutions should allow qualitative data sharers to choose their sharing strategies.

10.2.2 Institutions

The data curators in this dissertation study expressed their concern about the low awareness of data sharing in social sciences. This dissertation study also confirms that the data-sharing practices of early-career social scientists is unsatisfactory. However, it is still unclear what the root cause of this is, given that every stakeholder in the literature review (publishers, funders, professional associations) and all the participants in this dissertation study (early-career social scientists, social scientists who have data-sharing experience, employees working in a social science data repository) appear to be supportive of data sharing. As a bottom-up approach, an institution can act to engage early-career social scientists in the data sharing culture. In particular, participants in PS1 and CS3 expressed their expectations for their respective institutions, including a desire for strengthening technological support related to data sharing-reuse activities; in PS2, participants' perception of technological support is also positively associated with their data-sharing behaviors. To cultivate data sharing, institutions can engage and expose early-career social scientists (i.e., senior graduate students, post-doctoral researchers, and assistant professors) to training on data sharing preparation and to advocate data transparency, which is one of the foundations of open research (Lyon, 2016).

In addition, institutions should reconsider a reward system, as described in Section 9.4., such that the qualitative data sharing returns outweigh disclosure risks and time-consuming documentation work. To incentivize data sharing behaviors, institutions can consider data metrics and citations as an additional indicator for the promotion or recognition of faculty and researchers, since data sharing not only helps advance research but also serves the community.

10.2.3 Discipline communities

Since research norms and cultures are often discipline-specific, the best role for a discipline community is to provide a roadmap and guidelines for best practices. The study results of this dissertation further stress this importance, as participants in PS1 and PS3 strongly assert the need for establishing a best practice, one that can also be pushed forward by the discipline community. The discipline community can investigate discipline-level best practices, and professional associations can also provide data sharing guides. Such guides can not only help researchers prepare data sharing anonymization, but also prepare qualitative researchers to make informed decisions (e.g., which data type to share) when planning research.

As for discipline journals, journal editors should acknowledge alternatives to sharing raw data; that is, they should allow tools, coding results or selective interview transcripts to be shared as alternatives to full sets of interview transcripts.

10.2.4 Data repositories

Data is the key component in data repositories. Hence, data repositories have strong incentives to promote data sharing. However, as described in Chapter 8, data repositories are concerned about social scientists' low awareness about data sharing.

The data life cycle contains not only data storage but also data sharing and reuse. Hence, to advocate discipline sharing, data repositories should focus on data metrics, promote data reuse, and simplify data discovery.

On the other hand, data repositories can also advocate existing metadata standards such as DDI. For example, ICPSR provides online guidance and documentation on metadata standards. Moreover, except for QDR and ICPSR, there is little awareness about qualitative data sharing exemplars. Discipline data repositories can provide concrete examples of qualitative datasets, which will help researchers prepare their own qualitative data. These examples can be consulted when researchers are referring to the disciplinary best practice guide.

10.2.5 National policy makers

At the national level, policy makers can coordinate resources, create flexible policies, and study the balance between transparency and privacy.

The national government is in a position to coordinate different stakeholders (e.g., individuals, departments, institutions, discipline community associations, government, and research data infrastructures) and create a high-level roadmap to raise awareness of and develop best practices for qualitative data sharing. Raising awareness about data sharing requires contributions from all relevant stakeholders.

The data sharing mandates from social science-related national funders (such as NSF SBE and the Institute of Education Sciences (IES)) still adhere to STEM-like data sharing policies. This dissertation advocates: One policy cannot fit all disciplines. A national policy should examine existing mandates and policies to formulate flexible guidelines for social science data sharing, such that social scientists can explore the possible tradeoffs between data confidentiality and data transparency. Especially for qualitative data, policy makers should consider a "minimal standard" for sharing qualitative data, since sharing research tools or selective records is better than not sharing at all. Individual researchers are then encouraged to keep to the minimal standards, but try to follow the best practices.

The dissertation results also reveal the discrepancy between the definition of raw data by NSF and the definition of sharable data by social scientists who have experience sharing qualitative data. That is, social science-related funding agencies, such as the NSF SBE and the Institute of Education Sciences (IES), clearly address the importance of raw data; however, findings in this dissertation reveal that researchers are more willing to share tools than raw data.

In addition, policy makers should investigate the balance between privacy and transparency, and try to guide qualitative researchers toward a balanced strategy that can address both research transparency and confidentiality concerns. More concretely, policy makers should consider how to ensure full-disclosure and prevent disclosure risks during qualitative data sharing.

10.3 SUMMARY OF CONTRIBUTIONS

This dissertation study provides facts, insights, and guidance for social scientists, which helps facilitate data sharing and post-sharing curation in social sciences. While gaining more insight and understanding about individual researchers' data-sharing practices and infrastructural barriers, the instrument and research findings of this dissertation study can inform and contribute to several layers of stakeholders: individual, institutional, disciplinary community, data infrastructures, and national policy.

10.3.1 Individual layer

PIs who conduct research. Although previous work has mentioned challenges, concerns, motivations and benefits for sharing qualitative data, how those factors actually influence researchers' decisions and behaviors has not been sufficiently specified. This dissertation study, which identifies and examines cues that lead to qualitative scholars' data-sharing practices, is expected to help researchers who are interested in studying data archiving and sharing.

This dissertation also discusses strategies to develop the best practices of data sharing in social sciences. Such strategies can help qualitative researchers make better decisions about sharing their research data.

Based on participants' responses, this dissertation confirms that the lack of incentives is one major obstacle hindering data sharing. To motivate data sharing, one solution is to establish reward mechanisms, such as data citation. Moreover, the unique characteristics of qualitative data sharing (such as privacy concerns) demand more flexible policies to be adopted by the stakeholders (e.g., institutions and journal publishers).

Researchers and practitioners in digital curation fields. As for researchers who are interested in digital curation, the research framework, instrument, factual findings, and implications presented in this dissertation can serve as a foundation for further research studies. In particular, the proposed research framework and instrument can be applied to the investigation of data sharing and curation in other disciplines.

10.3.2 Institution layer- academic libraries and institutional repositories

In addition to researchers who are interested in data curation, this dissertation can also assist institutions that have need to serve and support researchers in digital curation.

One of the preliminary studies (Section 4.1) in this dissertation identifies the most developed areas and the least developed areas in terms of capability, and thus offers libraries or institutions a roadmap to prioritize the development of related services. Moreover, as qualitative researchers have been previously under-investigated, this dissertation's findings about qualitative researchers can help libraries and institutions navigate toward effective data services and consultations for qualitative researchers.

In addition to the abovementioned need, the instruments and experiences presented in this dissertation can also assist academic institutions that have a need to serve and support their researchers in digital curation (e.g., research data services or institution repositories). The instruments developed in this dissertation study to gather information about researchers' data activities and their perceptions about institutional supports can be used by local academic institutions to investigate their clients' needs and desires. The results can be used to build (or consider building) support for their faculty members, researchers, and students.

10.3.3 Discipline community layer

Academic domains e.g., the LIS community, political science community, and anthropology community. The discipline communities in social science domains can also benefit from this dissertation study because it identifies challenges and opportunities in terms of data sharing in social sciences. As more professional associations and academic communities are aware of the importance of data management, curation, and sharing, such an outcome can assist in the development of innovative toolkits and ethical guidelines in response to researchers' best practices.

Journal publishers. Based on the in-depth investigation of qualitative data sharing, this dissertation provides concrete suggestions, such as recognizing confidentiality concerns (informants' privacy and disclosure risks) encountered by qualitative researchers. Since these suggestions are derived based on empirical data, journal publishers can use them as solid references when making data-sharing policies, or adjust their current policies for qualitative studies.

10.3.4 Infrastructure layer – large-scale data infrastructures

This dissertation study can also contribute to data infrastructure. Within discipline data repositories, for example, instruments in this dissertation study can be used to investigate both data consumers' and data sharers' behaviors and practices. Data curation professionals can also benefit from the result findings by improving their understanding of social science researchers' barriers, perceived supports, and motivations to share data. In summary, the expected outcome can inform data repository staff how to re-frame or modify research data management services and resources, to reflect the infrastructural barriers and support structures that individual scholars perceive or experience.

10.3.5 National policies and global impacts

National layer. This dissertation investigates qualitative data-sharing practices in the U.S. It is envisioned that this dissertation can serve as an exemplar for other information professionals and researchers, and help national policy makers make informed decisions regarding qualitative data sharing in social sciences. For example, the findings highlighted in this dissertation suggest that policy makers should put more emphasis on norms, disclosure risks, and privacy, as the balance between privacy and transparency remains unclear and requires further study.

The NSF directorates related to social sciences, such as SBE, utilize the same data management policy as STEM disciplines (e.g., NSF Engineering Directorate, 2011). However, qualitative data has very different characteristics from its quantitative counterpart. While quantitative data deals with numerical values, qualitative data includes descriptions, concepts, and meanings mediated mainly through language and behaviors (Dey, 1993). Hence, developing universal guidelines to encourage data sharing might not reflect the different (and difficult) nature of qualitative data in social science disciplines. This dissertation study can provide evidence that qualitative and quantitative data are distinct by nature, and thus may require different management policies.

Global layer. Although European counties have established qualitative data archives for years, there are few studies based on empirical data. This dissertation attempts to fill this gap by performing extensive empirical studies. As this dissertation reflects the current situation of qualitative data sharing at the national level, its proposed research framework can be applied to performing health checks in countries outside the U.S as well.

10.4 LIMITATIONS

The instrument design, execution of research approaches, and sampling methods are the main limitations of this dissertation study.

10.4.1 Sampling approaches and sample size

The results might be biased due to the sampling approaches and sample size. The sampling approaches in CS1 and CS2 are based on convenience sampling (CS1) and voluntary responses (CS1 and CS2). Extra attention is required to interpret the generalizability of the results via convenience sampling. Considering the response rates in this dissertation study range from 11.8% to 16.8%, the voluntary approach used in CS1 and CS2 may tend to over-sample those who have relatively strong views or developed interest in the questionnaire theme (i.e., self-selection bias) and under-sample those who do not have interest in the topic (i.e., non-response bias). Therefore, selective bias is unavoidable, as is the case with all other social research using convenience sampling and voluntary responses. More specifically, there may be a bias toward people who are already aware of, have developed some interest in, or have strong opinions about data sharing in social sciences.

Another selection bias is caused by the sampling rationales. Since CS1 seeks out early-career researchers and CS2 targets experienced PIs, this dissertation may be biased toward a polarized sample: PhD students and full-ranked professors.

Finally, a sufficiently larger sample size in CS1 and CS2 would have allowed this dissertation to yield a more robust analysis outcome and "guard against overgeneralization" (Babbie, 2008, p7).

10.4.2 Self-administered survey

As with most of the surveys hosted on online platforms, Instrument 1 and Instrument 2 have limitations in this dissertation. Self-administration measures are known to have constraints on self-belief, and result in the under-reporting of behaviors that seem inappropriate, or responses that are perceived to be socially desirable (i.e., social desirability bias) (Donaldson, & Grant-Vallone, 2002). Therefore, although there is no existing literature or evidence to provide justification for the possible bias on researchers' data sharing behaviors, questions and responses involving moral judgements in this dissertation study, such as research integrity (protecting participants in CS1) and altruistic behaviors (in CS1 and CS2), should be interpreted with caution.

10.4.3 Data triangulation

The data triangulation in this dissertation study may amplify the qualitative part of data collection. The three case studies in this dissertation are led by individual aspects of two central research questions. However, due to the limitation of the instrument design and its derived data collection, this dissertation study may have a selection bias toward qualitative data in the triangulation process. The reason is that individual open-ended responses in CS1 and CS2 are more informative and easily comparable to CS3 results. Participants' comments in the open-ended questions in CS1 and CS2 may be easily mentioned or quoted as evidence during the data triangulation.

10.5 DIRECTIONS FOR FUTURE WORK

While this dissertation study has investigated qualitative data-sharing practices in social sciences, there are opportunities for extending the research scope of this study. This section presents some of these directions.

One of the major directions is to extend the discipline scope to behavioral science or humanities, and even to the qualitative studies in health sciences, to test if the profile tool in CS1 is well-adopted or the survey in CS2 can be generalized to other disciplines.

Second, the results have triggered two more points of interest.

- What is the tension or how can a balance be struck between research transparency and concerns about confidentiality regarding qualitative data sharing? How can full-disclosure be ensured, and how can disclosure risks be avoided during qualitative data sharing?
- What role can different stakeholders (individuals, departments, institutions, discipline community associations, government, and research data infrastructures) play in raising the awareness of and developing the best practices for qualitative data sharing?

Finally, based on this dissertation study, technologies used for social science research are very dispersed, and this phenomenon reflects diverse research inquiries and their approaches. One possible future direction is to develop a tool or service to value and preserve social research methods and heterogeneous data. One entry point can be conducting participatory action research that engages stakeholders in the data sharing-reuse process, to invite them to participate and design a prototype that can support their workflow in the data sharing-reuse process.

10.6 CONCLUSION

This dissertation examines qualitative data-sharing practices in social sciences, which have been thus far under-investigated by related work.

By synergizing the theory of Knowledge Infrastructure (KI) and the Theory of Remote Scientific Collaboration (TORSC), this dissertation study develops a series of instruments to investigate data-sharing practices in social sciences. Two preliminary studies and three case studies are conducted and triangulated to answer the inquiry in four dimensions of the topic: data characteristics, individual, technological, and organizational aspects.

The triangulation of all studies in this dissertation further unveils several important findings about data sharing in social sciences, including:

- Data aspects: The confusion about data ownership and its research value should be addressed before researchers can confidently share data. In addition, when it comes to sharable qualitative data, most researchers think about sharing research tools but not the actual data from the informants. Therefore, this dissertation study suggests that funders or institutions should consider different data sharing granularities, thereby allowing qualitative data sharers to choose their sharing strategies from full disclosure, partial disclosure, or minimum standards for data regarding potentially vulnerable individuals.
- Organizational context: To foster data sharing, the community plays a key role to catalyze the development of best practices of sharing data.
- Individual motivation: Since social scientists who have data-sharing experience often seek concrete reward such as citations or career promotion, the discipline community and institution should consider providing incentives in such fashion.

- Technological supports: Despite the maturity of DDI and ICPSR endeavors, the majority of the social scientists were unware of the standard and procedures of data sharing. Moreover, they believe that ideal technology should enable seamless workflow and support management of various research products.

This dissertation study seeks to pave the way for understanding the contemporary research infrastructure in social sciences based on empirical data collection. The results and triangulation among sub-studies provide strategies to the best practices of data sharing in social sciences. The implications can inform current decisions, guidelines, and policies which can craft a more sustainable data-sharing environment in social sciences and beyond.

11.0 BIBLIOGRAPHY

Abbott, A. (2001). *Chaos of Disciplines*. Chicago, IL: University of Chicago Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211.

American Anthropological Association. "AAA." (2009). AAA code of ethics. AAA. Retrieve from http://www.americananthro.org/

APSA (2012). A guide to professional ethics in political science. American Political Science Association. Retrieved from http://www.apsanet.org/portals/54/Files/Publications/APSAEthicsGuide2012.pdf

Archer, T. M. (2008). Response rates to expect from Web-based surveys and what to do about it. *Journal of Extension*, 46(3) Article 3RIB3. Retrieved from https://www.joe.org/joe/2008june/rb3.php

ARL (n.d.). E-Research. Retrieved April 9, 2015 from http://www.arl.org/focus-areas/e-research#.VSb8aWjF-rk

Australian Data Archive. "ADA." (n.d.). Retrieved from https://www.ada.edu.au/

Babbie, E. R. (2008). *The Practice of Social Research*. Belmont, CA: Wadsworth Publishing Company.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66(3), 411-421.

Barbour, R. (2007). *Introducing Qualitative Research: A Student's Guide to the Craft of Doing Qualitative Research*. London: Sage.

Beecher, B. (2009). The ICPSR pipeline process. Retrieved from http://techaticpsr.blogspot.com/2009/11/icpsr-pipeline-process.html

Bhattacherjee, A. (2012). *Social Science Research: Principles, Methods, and Practices*. University of South Florida Tampa Bay Open Access Textbooks Collection Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=oa_textbooks

Bishop, L. (2005). Protecting respondents and enabling data sharing: Reply to Parry and Mauthner. *Sociology*, 39(2), 333-336.

Bishop, L. (2007). A reflexive account of reusing qualitative data: Beyond primary/secondary dualism. *Sociological Research Online*, 12(3), 2.

Bishop, L. (2009). Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues,* 44(3): 255–272.

Bishop, L. (2016). Sharing qualitative data: Challenges and opportunities. *University of Central Lancashire Open Scholarship Month Event*. Retrieved from https://www.ukdataservice.ac.uk/media/604298/bishop_qualdatasharing_ucentrallanc_2march2016.pdf

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly,* (3)4. Retrieve from http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press.

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The ups and downs of knowledge infrastructures in science: implications for data management. *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, 257-266.

Borreani, C., Miccinesi, G., Brunelli, C., & Lina, M. (2004). An increasing number of qualitative research papers in oncology and palliative care: does it mean a thorough development of the methodology of research. *Health Qual Life Outcomes*, 2, 7.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research*. Dordrecht: Springer Netherlands.

Bowler, L., Knobel, C., & Mattern, E. (2015). From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology*, 66(6), 1274-1293.

Broom, A., Cheshire, L., & Emmison, M. (2009). Qualitative researchers' understandings of their practice and the implications for data archiving and sharing. *Sociology*, 43(6), 1163–1180.

Buckingham, D. (2009). "Creative" visual methods in media research: possibilities, problems and proposals. *Media, Culture & Society*, 31, 4-652.

Cheshire, L. (2009). Archiving qualitative data: Prospects and challenges of data preservation and sharing among Australian qualitative researchers. The Australian Qualitative Archive (AQuA). Retrieve from http://www.assda.edu.au/forms/AQuAQualitativeArchiving_DiscussionPaper_FinalNov09.pdf

Cliggett, L. (2013). Qualitative data archiving in the digital age: strategies for data preservation. *The Qualitative Report*, 18(24).

Clubb, J. M., Austin, E. W., Geda, C. L., & Traugott, M. W. (1985). Sharing research data in the social sciences. In S. E. Fienber, M.E. Martin, & M. L. Straff (Eds.), *Sharing Research Data*, (39-88). DC: National Academies Press.

Connelly, F. M., & Clandinin, D. J. (1990). Stories of experience and narrative inquiry. *Educational Researcher*, 19(5), 2-14.

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and Sharing Research Data: A Guide to Good Practi*ce. London: Sage.

Council on Library and Information Resources. (2013). Research Data Management Principles, Practices, and Prospects. CLIR. Retrieve from https://www.clir.org/pubs/reports/pub160/pub160.pdf

Costas, R., Meijer, I., Zahedi, Z. and Wouters, P. (2013). The value of research data - metrics for datasets from a cultural and technical point of view. *A Knowledge Exchange Report*, Retrieve from http://www.knowledge-exchange.info/datametrics

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions*. Series A, Mathematical, Physical, and Engineering Sciences, 368(1926), 4023–4038.

Creswell, J. (2009). Research design: Qualitative, quantitative, and mixed methods approaches. Sage.

Creswell, J. (2013). Qualitative, quantitative, and mixed methods approaches. In Research Design (pp. 1–26). Sage

Curty, R, G. (2016). Factors influencing research data reuse in the social sciences: an exploratory study. *International Journal of Digital Curation*, 11(1): 96-117.

Curty, R. G; Kim, Y.; and Qin, J. (2013). What have scientists planned for data sharing and reuse? a content analysis of NSF awardees' data management plans. *iSchool Post-doc and Student Scholarship*. Paper 2. http://surface.syr.edu/ischoolstudents/2

Curty, R. G., & Qin, J. (2014). Towards a model for research data reuse behavior. *Proceedings of the Association for Information Science and Technology Annual Meeting*, 51(1).

Data Curation Profiles (n.d.). Retrieved from http://datacurationprofiles.org/

de Montalvo, U. W. (2003). In search of rigorous models for policy oriented research: A behavioural approach to spatial data sharing. *URISA Journal*, 15(1), 19-28. Chicago

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Springer Science & Business Media.

Dey, I. (1993). Qualitative data analysis: A user friendly guide for social scientists. Routledge.

Diekema, A. R., Wesolek, A., & Walters, C. D. (2014). The NSF/NIH Effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. *Journal of Academic Librarianship*, 40(3-4), 322-331.

Digital Curation Centre ("DCC"). (2017). Disciplinary Metadata. Retrieved from http://www.dcc.ac.uk/resources/metadata-standards

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Bowker, C. L., Ribes, D., Burton,. M., & Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenge*s. Retrieved from http://pne.people.si.umich.edu/PDF/Edwards_etal_2013_Knowledge_Infrastructures.pdf

Elman, C., & Kapiszewski, D. (2013). *A Guide to Sharing Qualitative Data. Center for Qualitative and Multi Method Inquiry (CQMI)*, Syracuse University.

Elman, C., and Kapiszewski, D. (2014). Data access and research transparency in the qualitative tradition. *Political Science and Politics*, 47(1): 43–47.

Elman, C., Kapiszewski, D., & Vinuela, L. (2010). Qualitative data archiving: Rewards and challenges. PS: Political Science & Politics, 43(01), 23-27.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statisti*cs, 5(1), 1-4.

Faisal, (2008). *Academic Thesis Generation*. Retrieved from https://drfaisallearntips.wordpress.com/page/2/

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing?. *PLOS ONE*, 10(2), e0118053.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). *Sharing Research Data.* National Academies Press.

Fink, A. S. (2000). The role of the researcher in the qualitative research process. A potential barrier to archiving qualitative data. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 1(3).

Fraenkel, J. R. & Wallen, N. E. (2003). *How to Design and Evaluate Research in Education* (5th ed.). Boston: McGraw-Hill.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, 4(4), 540–553.

Friedlander, A. (2009). Asking questions and building a research agenda for digital scholarship. *Working Together or Apart: Promoting The Next Generation of Digital Scholarship*, 1-15.

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J. and Rasmussen, B. (2009). Identifying Benefits arising from the Curation and Open Sharing of Research Data. UK Higher Education and Research Institutes, November. Retrieved from http://ie-repository.jisc.ac.uk/279/

Gagné, M. (2009). A model of knowledge-sharing motivation. *Human Resource Management*, 48(4), 571-589.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2), 349-360.

Gliem, R. R., & Gliem, J. A. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.

Goben, A., & Salo, D. (2013). Federal research data requirements set to change. *College & Research Libraries News*, 74(8), 421-425.

Gómez, C. C. (2009). Assessing the Quality of Qualitative Health Research: Criteria, process. and writing. *Forum: Qualitative Social Research*, 10(2), 1–19.

Griffin S. (2015). Libraries in the digital age: technologies, innovation, shared resources and new responsibilities, In L. Cantoni & J. Danowski (Eds), *Communication and Technology, Volume 5 of the series Handbook of Communication Science*. De Gruyter Mouton.

Guest, G., Namey, E. E., & Mitchell, M. L. (2012). *Collecting Qualitative Data: A field manual for applied research*. Sage.

Gutmann, M. P., Evans, B., Mitchell, D., & Schürer, K. (2009). The Data Archive Technologies Alliance: Looking towards a Common Future. *Annual Meeting of the International Association for Information Service and Technology (IASSIST)*. Tampere, Finland.

Haggerty, K. D. (2004). Ethics creep: Governing social science research in the name of ethics. *Qualitative Sociology*, 27(4), 391-414.

Halbert, M. (2013). *Prospects for Research Data Management*. CLIR Publication,160. Retrieved from http://www.clir.org/pubs/reports/pub160/pub160.pdf

Hammersley, M. (1997). Qualitative data archiving: some reflections on its prospects and problems. *Sociology*, 31(1), 131-142.

Hayes, F. (1997). Research Methods and Statistics: Lecture and Commentary Notes. Retrieved from http://webstat.une.edu.au/unit_materials/index.htm

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends* 57 (2): p. 280-299.

Hey, A. J., & Trefethen, A. E. (2003). The data deluge: An e-science perspective. Grid Computing – Making the Global Infrastructure a Reality. Retrieved from http://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research.

Holliday, A. (2007). *Doing & Writing Qualitative Research*. Sage.

IASSIST (n.d.). IASSIST Home . Retrieved from http://www.iassistdata.org/

ICPSR. (2016). Size of ICPSR's Holdings. Retrieved October 31, 2016, from https://www.icpsr.umich.edu/icpsrweb/content/about/history/

ICPSR. ICPSR: A Case Study in Repository Management. Retrieve from https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html

IES. (n.d.). Data Sharing Implementation Guide. Retrieved April 17, 2015, from http://ies.ed.gov/funding/datasharing_implementation.asp

Inter-university Consortium for Political and Social Research ("ICPSR") (2010). Preparing data for sharing; Guide to social science data archiving. *Data Archiving and Networked Services – DANS. Pallas Publications* Amsterdam: Amsterdam University Press.

Inter-university Consortium for Political and Social Research ("ICPSR") (2012), Guide to Social Science Data Preparation and Archiving, 5th ed., ICPSR, Ann Arbor, MI, available at: www.icpsr.umich.edu/access/dataprep.pdf

Inter-university Consortium for Political and Social Research ("ICPSR") (2014.). Grants and Contracts Fiscal Year 2013. Retrieved July 31, 2014, from http://www.icpsr.umich.edu/icpsrweb/content/membership/grants.html

Inter-university Consortium for Political and Social Research ("ICPSR") (n.d.). Retrieved from http://www.icpsr.umich.edu/

Israel, M. (2015). *Research Ethics and Integrity for Social Scientists: Beyond Regulatory Compliance*. Sage.

Israel, M., & Hay, I. (2006). Research ethics for social scientists. Sage.

Jahnke, L., Asher, A., & Keralis, S. D. C. (2012). *The Problem of Data*. DC: Council on Library and Information Resources.

Jansen, H. (2010). The logic of qualitative survey research and its position in the field of social research methods. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Vol. 11, No. 2).

Jeng, W. & Lyon, L. (2016). A report of data-intensive capability, institutional support, and data management practices in social sciences. *International Journal of Digital Curation*, 11(1): 156-171.

Johnson, B., & Christensen, L. (2008). *Educational Research: Quantitative, Qualitative, and Mixed Approache*s. Sage.

Johnson, W. G. (2008). The ICPSR and social science research. *Behavioral & Social Sciences Librarian*, 27(3-4), 140-157.

Karcher, S., Kirilova, D., & Weber, N. (2016). Beyond the matrix: Repository services for qualitative data. Moynihan Institute of Global Affairs. Paper 1. Retrieved from http://surface.syr.edu/miga/1/

Kanfer, A. G., Haythornthwaite, C., Bruce, B. C., Bowker, G. C., Burbules, N. C., Porac, J. F., & Wade, J. (2000). Modeling distributed knowledge processes in next generation multidisciplinary alliances. *Information Systems Frontie*rs, 2(3-4), 317-331.

Kanfer, R., Chen, G., & Pritchard, R. D. (Eds.) (2008). *Work Motivation: Past, Present, and Future*. New York: Taylor and Francis Group.

Kim, Y. (2013). *Institutional and Individual Influences on Scientists' Data Sharing Behaviors*. Unpublished dissertation. Syracuse University.

Kim, Y., & Stanton, J. M. (2012). Institutional and individual influences on scientists' data sharing practices. *Journal of Computational Science Education*, 3(1), 47.

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, 28(03), 444-452.

Kjeldgaard, A. S. F. (2010). Archiving and disseminating qualitative data in Denmark. *IASSIST Quarterly*, 34(3/4), 35.

Krauwer, S., & Hinrichs, E. (2014). The CLARIN research infrastructure: resources and tools for e-humanities scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Krawczyk, M., & Reuben, E. (2012). (Un)available upon request: field experiment on researchers' willingness to share supplementary materials. *Accountability in Research*, 19(3), 175-186.

Kuipers, T., & Hoeven, J. (2009). Insight into digital preservation of research output in Europe. PARSE Survey Report.

Kuo. (2011). Mixed research and the qualitative quantitative debate. *Soochow Journal of Political Science,* 29 (1), 1-64.

Kuula, A. (2011). Methodological and ethical dilemmas of archiving qualitative data. *IASSIST Quarterly*, 34(3/4), 35.

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder. portal: *Libraries and the Academy*, 11(4), 915-937.

Latham, G. P., & Pinder, C. C. (2005). Work motivation theory and research at the dawn of the twenty-first century. *Annual Review of Psychology*, 56, 485-516.

Lavoie, B. F. (2004). The open archival information system reference model: Introductory guide. *Microform & Imaging Review*, 33(2), 68-81.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science*, 323(5915), 721.

Leland Speed Library at Mississippi College (n.d.) Research Process - Subject Guides - LibGuides. Retrieved from http://mc.libguides.com/eddoc/research

Lim, H. B., Iqbal, M., Yao, Y., & Wang, W. (2010). A smart e-Science cyberinfrastructure for cross-disciplinary scientific collaborations. *Semantic e-Science*, 67-97. Springer.

Lin, H. F. (2007). Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. *Journal of Information Science*, 33 (2), 135–149.

Lupia, A., & Elman, C. (2014). Openness in Political Science: Data Access and Research Transparency. *PS: Political Science & Politics*, 47(01), 19-42.

Lyon, L., Ball, A., Duke, M., & Day, M. (2012). *Community Capability Model Framework*. Retrieved from http://communitymodel.sharepoint.com/Pages/default.aspx

Lyon, L., Patel, M., & Takeda, K. (2014). Assessing requirements for research data management support in academic libraries: introducing a new multi-faceted capability tool. *Libraries in the Digital Age (LIDA) Proceedings*, 13.

Lyon, L. (2016). Transparency: the emerging third dimension of Open Science and Open Data. Liber quarterly, 25(4).

Lyon, L., Jeng, W., & Mattern, E. (2017). Research transparency: A preliminary study of disciplinary conceptualisation, drivers, tools and support services. To appear in 12th International Digital Curation Conference.

Mackey, K. (2009) Research Process. Retrieved from http://www.clark.edu/Library/iris/types/research_process/research_process_p3.shtml.

Malins, J., & Gray, C. (2013). *Visualizing Research: A Guide to the Research Process in Art and Design*. Ashgate Publishing, Ltd.

Mark & Helen Osterlin Library. (n.d.) Seven Steps of Research: A Map of the Research Process.

Northwestern Michigan College. Retrieved August 24, 2015 from http://web.csulb.edu/~ttravis/test/researchmap.html

Martin, V. (2014). *Demystifying eResearch: A Primer for Librarians*. Santa Barbara, CA: Libraries Unlimited.

Mason, J. (2007). 'Re-Using' qualitative data: on the merits of an investigative epistemology. *Sociological Research Online*, 12(3), 3.

Mattern, E, Jeng, W., He, D., Lyon, L., & Brenner, A. (2015). Using participatory design and visual narrative inquiry to investigate researchers' data challenges and recommendations for library research data services. *Program: Electronic Library and Information Systems*. 49(4): 408-423.

Mauthner, N. S., & Parry, O. (2009). Qualitative data preservation and sharing in the social sciences: On whose philosophical terms? *Australian Journal of Social Issues*, 44(3), 289-305.

Mauthner, N. S., Parry, O., & Backett-Milburn, K. (1998). The data are out there, or are they? Implications for archiving and revisiting qualitative data. *Sociology*, 32(4), 733-745.

Maynard, D. W., & Schaeffer, N. C. (2000). Toward a sociology of social scientific knowledge survey research and ethnomethodology's asymmetric alternates. *Social Studies of Science*, 30(3), 323-370.

Mennes, M., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2013). Making data sharing work: The FCP/INDI experience. *Neuroimage*, 82, 683-691.

Mischo, W. H., Schlembach, M. C., & O'Donnell, M. N. (2014). An analysis of data management plans in university of illinois national science foundation grant proposals. *Journal of eScience Librarianship*, 3(1), 3.

Moore, N. (2007). (Re)Using Qualitative Data? *Sociological Research Online*, 12 (3), 1.

Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry*, 1077800413513733.

Mori, H., & Nakayama, T. (2013). Academic impact of qualitative studies in healthcare: bibliometric analysis. *PLOS ONE*, 8(3), e57371.

Morrow, V., Boddy, J., & Lamb, R. (2014). The ethics of secondary data analysis: learning from the experience of sharing qualitative data from young people and their families in an international study of childhood poverty.

Motivation theory. (2009). In L. Sullivan (Ed.), *The SAGE Glossary of the Social and Behavioral Sciences.* (pp. 333-334). Thousand Oaks, CA: Sage.

Myers, J., Hedstrom, M., Akmon, D., Payette, S., Plale, B. A., Kouper, I., ... & Kumar, P. (2015). Towards sustainable curation and preservation: The sead project's data services approach. In e-Science (e-Science), 2015 IEEE 11th International Conference on (pp. 485-494).

National Science Foundation (NSF). (2013). National Science Foundation's Merit Review Criteria: Review and Revisions. Retrieved October 31, 2016, from https://nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_sigchanges.jsp

National Science Foundation. (2014). Survey of Earned Doctorates (SED). Retrieved December 26, 2016, from http://www.nsf.gov/statistics/srvydoctorates/

NEH. (n.d.). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by the National Endowment for the Humanities. Retrieved from http://www.neh.gov/about/guidelines-for-information-disseminated-by-national-endowment-for-humanities

NSF Data Sharing Policy (n.d.). Dissemination and Sharing of Research Results. Retrieved from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301-314.

O'Carroll, A. (2011). Qualitative research in Ireland. *IASSIST Quarterly*, 19.

Office of Science and Technology Policy. (2000). Federal research misconduct policy. Federal Register, 65(235), 76260-76264.

Olson, G. M., & Olson, J. S. (2000). Distance matters. Human-computer interaction, 15(2), 139-17.

Olson, G. M., Zimmerman, A., & Bos, N. (2008). Scientific collaboration on the Internet. Cambridge, MA: MIT Press.

Olson, J. S., & Olson, G. M. (2013). Working together apart: Collaboration over the internet. Synthesis Lectures on Human-Centered

Papastatidis, S. (2009). A Platform for All that We Know: Creating a Knowledge-Driven Research Infrastructure. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. 165–172). Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx

Parry, O., & Mauthner, N. S. (2004). Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology*, 38(1), 139-152.

Parry, O., & Mauthner, N. (2005). Back to basics: Who reuses qualitative data and why?. *Sociology*, 39(2), 337-342.

Parsons, M. A., Godøy, Ø., LeDrew, E., De Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37(6), 555-569.

Patton, M. (2001). *Qualitative Research & Evaluation Methods*. Sage.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373-1379.

Peterson, E. R., & Barron, K. A. (2007). How to get focus groups talking: New ideas that will stick. *International Journal of Qualitative Methods*, 6(3), 140-144.

Pink, S. (2006). *The Future of Visual Anthropology: Engaging the Senses*. Taylor & Francis.

Poline, J.B., Breeze, J.L., Ghosh, S.S., Gorgolewski, K., Halchenko, Y.O., Hanke, M., Helmer, K.G., Marcus, D.S., Poldrack, R.A., Schwartz, Y. and Ashburner, J. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6, 9.

Prescott, A. (2013). Big Data Requirements in Arts and Humanities. [PowerPoint slides]. Retrieved from http://indico.cern.ch/event/246453/session/4/contribution/35/material/slides/1.pdf

Qualidata, E. S. D. S. (2012). About ESDS Qualidata. Universities of Essex and Manchester. Retrieved from http://www. esds. ac. uk/qualidata/about/introduction. asp.

Qualitative Data Model Working Group. "QDMWG" (n.d.). Retrieved April 9, 2015, from http://www.ddialliance.org/alliance/working-groups#qdewg

Qualitative Data Repository. "QDR" (n.d.). Retrieved from https://qdr.syr.edu/

Rasmussen, K. B. (2011). Barking up the right tree. Editor's notes. IASSIST Quarterly.

Ratner, C. (2002). Subjectivity and objectivity in qualitative methodology. *In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 3(3).

Reeve, J. (2001). *Understanding Motivation and Emotion*. New York: Wiley.

Resnik, D. B. (2010). What is ethics in research & why is it important. Research Triangle Park, North Carolina: National Institute of Environmental Health Sciences/National Institute of Health.

Responsible Conduct in Data Management - The Office of Research. (n.d.). Retrieved from https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html

Ribes, D. and T. A. Finholt (2009). The long now of infrastructure: Articulating tensions in development. *Journal for the Association of Information Systems*, 10(5): 375-398.

Richards, L. (2014). Handling qualitative data: A practical guide. Sage.

ROARMAP: Registry of Open Access Repositories Mandatory Archiving Policies. (n.d.). Retrieved November 7, 2014, from http://roarmap.eprints.org/

Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods. In Denzin & Lincoln (Eds.) *Handbook of qualitative research.* (2nd Edition). 769-802.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68.

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, 30, S19-S31.

Sekaran, U. (2006). *Research Methods for Business: A Skill Building Approach*. John Wiley & Sons.

Sieber, J. E. (Ed.). (1991). *Sharing Social Science data: Advantages and Challenges* (Vol. 128). Sage.

Slavnic, Z. (2011). Preservation and Sharing of Qualitative Data-Academic Debate and Policy Developments. TheMES on Ethnic Studies. Linköping: REMESO.

Slavnic, Z. (2013). Towards qualitative data preservation and re-use—Policy trends and academic controversies in UK and Sweden. *In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 14(2).

Smioski, A. (2011a). Establishing a qualitative data archive in Austria. *IASSIST Quarterly*, 31.

Smioski, A. (2011b). Archiving qualitative data: Infrastructure, acquisition, documentation, distribution. Experiences from WISDOM, the Austrian Data Archive. *In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 12(3).

Social Science [Def. 1]. (n.d.). In Merriam Webster Online, Retrieved November 11, 2014, from http://www.merriam-webster.com/dictionary/social%20science

Social Science [Def. 1]. (n.d.). In Oxford Dictionaries, Retrieved November 11, 2014, from http://www.oxforddictionaries.com/us/definition/american_english/social-science

Sung, Y. T., & Pan, P. Y. (2010). Applications of mixed methods research in educational studies. *Journal of Research in Education Sci*ences, 55(4), 97-130.

Tancheva, K. (2012). "Linguistics - Cornell University," Data Curation Profiles Directory: Vol. 4, Article 7. DOI: http://dx.doi.org/10.5703/1288284315007

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling a typology with examples. *Journal of Mixed Methods Research*, 1(1), 77-100.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6).

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS ON*E, 10(8), e0134826.

Tong, A., Winkelmayer, W. C., & Craig, J. C. (2014). Qualitative Research in CKD: An Overview of Methods and Applications. *American Journal of Kidney Diseases*.

Tsai, A. C., Kohrt, B. A., Matthews, L. T., Betancourt, T. S., Lee, J. K., Papachristos, A. V., ... & Dworkin, S. L. (2016). Promises and pitfalls of data sharing in qualitative research. *Social Science & Medicine*, 169, 191-198.

Tsang, K. K. (2012). The use of midpoint on Likert Scale: The implications for educational research. *Hong Kong Teachers' Centre Journal*, 11, 121-130.

UK Data Service (n.d.). Research Data Lifecycle. Retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle

UK Data Service (n.d.). Reusing qualitative data. Retrieved April 3, 2015, from http://ukdataservice.ac.uk/use-data/guides/methods-software/qualitative-reuse.aspx

University of California Irvine Libraries (n.d.). Digital Scholarship Services. Retrieved from http://www.lib.uci.edu/dss/

University of California Museum of Paleontology (2008). How Science Works. Retrieved from http://undsci.berkeley.edu/lessons/pdfs/complex_flow_handout.pdf

University of Virginia Library Research Data Services (n.d.). Steps in the Data Lifecycle. Retrieved from http://data.library.virginia.edu/data-management/lifecycle/

University of Virginia Library Research Data Services. (n.d.). Retrieved April 2, 2015, from http://data.library.virginia.edu/data-management/plan/format-types/

Unsworth, J. (2006). Our Cultural Commonwealth: the report of the American Council of learned societies commission on cyberinfrastructure for the humanities and social sciences. ACLS: New York.

Vagias, Wade M. (2006). Likert-type scale response anchors. Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University.

Van den Eynden, V. and Bishop, L. (2014). Sowing the Seed: Incentives and motivations for sharing research data, a researcher's perspective. Retrieved from http://www.data-archive.ac.uk/media/492924/ke_report-incentives-for-sharing-research-data.pdf

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. Archival Science, 7(1), 73-87.

Viktor, Prokopenya (2008). Systems Model of Action-Research Process. Retrieved from commons.wikimedia.org/wiki/File:Systems_Model_of_Action-Research_Process.jpg

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS One,* 8(7), e67332.

White, D. (1991). Sharing anthropological data with peers and third world hosts. In J. Sieber (Ed.), SAGE Focus Edition: Sharing social science data: Advantages and challenges. (pp. 42-61). Thousand Oaks, CA: SAGE Publications, Inc.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE*, 6(11), e26828.

Williams, M., Dicks, B., Coffey, A., & Mason, B. (2007). Qualitative data archiving and reuse: mapping the ethical terrain. Methodological issues in qualitative data sharing and archiving.

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curatio*n, 4(3), 93-103.

Wolski, M., & Richardson, J. (2014). A Model for Institutional infrastructure to support digital scholarship. P*ublications*, 2(4), 83-99.

Wutich, A., & Bernard, H. R. (2016). Sharing qualitative data & analysis. With whom and how widely?: A response to 'Promises and pitfalls of data sharing in qualitative research'. *Social Science & Medicine*, 169, 199-200.

Yoon, A. (2016). Data reusers' trust development. *Journal of the Association for Information Science and Technology*. Early View. doi: 10.1002/asi.23730

Yoon, A., & Tibbo, H. (2011). Examination of Data Deposit Practices in Repositories with the OAIS Model. *IASSIST Quarterly*, 35(4).

Yoon, A., Hall, M., & Hill, C. (2014). "Making a square fit into a circle": Researchers ' experiences reusing qualitative data. *Proceedings of the American Society for Information Science and Technology*, 51(1).

Zilinski, L. & Lorenz, S. (2012). Linguistics / Etymology - University of South Florida. *Data Curation Profiles Directory*, 4, 9.

APPENDIX A. QUALITATIVE DATA TYPES (QDR)

Qualitative data types suggested by Qualitative Data Repository (QDR)

- Data from interviews; focus groups; oral histories (audio/video recordings; transcripts; notes/summaries; questionnaires/interview protocols)
- Field notes (including from participant observation or ethnography)
- Maps/satellite imagery/geographic data
- Official/public documents, files, reports (diplomatic, public policy, propaganda, etc.)
- Meeting minutes
- Government statistics
- Correspondence, memoranda, communiqués, queries, complaints
- Parliamentary/legislative proceedings
- Testimony in public hearings
- Speeches, press conferences
- Military records
- Court records; legal documents (charts, wills, contracts)
- Chronicles, autobiographies, memoirs, travel logs, diaries
- Brochures, posters, flyers
- Press releases, newsletters, annual reports
- Records, papers, directories
- Internal memos, reports, meeting minutes
- Position/advocacy papers, mission statements
- Party platforms
- Personal documents (letters, personal diaries, correspondence, personal papers)
- Maps, diagrams, drawings
- Radio broadcasts (audio or transcripts)
- TV programs (video or transcripts)
- Print media (magazine, newspaper articles)
- Electronic media
- Published collections of documents, yearbooks, etc.
- Books, articles, dissertations, working papers

- Photographs
- Ephemera; popular culture visual or audio materials (printed cloth, art, music /songs, etc.)

APPENDIX B. CUSTOMIZED CCMF INSTRUMENT (ANTHROPOLOGY)

About Your Research Data (open-ended questions)

- What is the subject discipline or sub-discipline to which your data relates?
- What types of data do you work on? e.g. observational, survey, experimental, reference, records, historical materials etc.
- What is the nature, range and scope of your research data? e.g. environmental, geographical, medical, astronomy, human behavioral, demographic etc.
- Can your data be recollected or recreated?
- Are your data sensitive or have ethical issues associated with them?
- What are the typical data volumes that you work with for one project (e.g., 10 gigabytes)?
- Do you consider your research as a data-intensive or compute-intensive one? ("Data-intensive" research is research that involves large amounts of data, possibly combined from many sources across multiple disciplines, and requires some degree of computational analysis. If research involves combining data from several different sources, where the different source datasets have been collected according to different principles, methods and models, and for a primary purpose other than the current one, then it is likely to be classed as data-intensive research.)
- How complex is your data? Does it contain multiple variables (attributes)? Please describe how complex you think they are. (e.g. inter-relationships with other datasets, both quantitative and qualitative data are collected)
- Have you used any tool (e.g., from library resources or a checklist) to assess your Research Data Management needs?
- What is the source of funding for your research and associated data?
- Have you ever used data that were not generated by you or your own team?
- Have you ever shared data with others?
- Do you ever put your data in an institutional repository or a data center? (i.e., an online data center for collecting, preserving, and disseminating digital copies of research datasets) If so, could you please list out?
- Have you ever accessed others' data in an institutional repository or a data center?

- Are you willing to share your data to others when receiving a request?
- Has anyone ever asked you to share your data? Please describe your experience.

1. Collaboration

Instruction: For this set of elements, consider the discipline to be the general area of science as well as specialization areas. To what extent to you engage in:

Data Table 1. CCMF- Collaboration items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
1.1 Collaboration within the discipline (e.g., anthropology or cultural studies)	None or Lone researchers.	Departmental research groups.	Collaboration across research groups within or between organizations.	Discipline organized at a national level.	International collaboration and consortia.
1.2 Collaboration and interaction across disciplines	None or limited	Individual researchers occasionally collaborate outside their discipline.	Disciplines collaborate through joint conferences or publications.	Bilateral collaborations.	Formal collaboration between research groups from several different disciplines.
1.3 Collaboration and interaction across sectors (e.g. public, private, government)	None or limited	Attempts have been made but are not considered successful.	Despite successful examples working with other sectors is not the norm – some barriers are perceived.	A discipline or group has gained experience of working closely with one or two sectors.	Work successfully with several other sectors on different problems
1.4 Collaboration with the public (e.g., engaging citizens)	None or limited	The public's involvement is limited to acting as subjects of study, user testing, etc.	Contact with the public is only through occasional appearance in the media e.g. news bulletins, TV programs	Mainly informational, sometimes participative, targeted media programs are organized to engage the public e.g. science fairs	Dedicated programs involving the public in research; Crowd sourcing/citizen science

2. Skills & Training

Instruction: For this set of elements, consider the extent to which training in data-related tools, techniques and issues is available to you as a researcher. To what degree are you aware of training in the following aspects. If you know of specific training, workshops, or tools, etc. please provide it on the comment section.

Data Table 2. CCMF- Skills and training items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
2.1 Research data management e.g. Use of tools for managing research data	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.2 Data Collection, Processing and Analysis (including management of private and sensitive data)	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.3 Data description and identification e.g. metadata schemes, controlled vocabularies such as AFS's Ethnographic Thesaurus for folklore datasets, digital identifiers (unique control	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.

identifiers of your data)					
2.4 Copyright and data licenses e.g. Creative Commons	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.5 Quality control, security, validity and integrity	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.6 Publication and sharing of research data (including human and automated processing)	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.7 Linking publications to research data	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.8 Making research data discoverable	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.

2.9 Finding, retrieving and repurposing existing datasets	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.10 Making research data reusable	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.11 Data referencing and data citation e.g. it uniquely identifies an object or a file stored in a repository	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.12 The concepts of measuring scholarly impacts on data e.g. Impact factor indexes of research datasets, alternative metrics of datasets such as the number of downloads or social media mentions	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.13 Management of research information and use of a research discovery/networking system e.g. Common Research	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing

Information System (CRIS)			mentoring on data management.	provided on request.	professional development.
2.14 Policy and planning e.g. data management, business models	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.
2.15 Collaboration (e.g., engaging with other researchers) and communication (e.g., engaging with the public or the media)	None or unknown	Training programs in development.	Training available but not embedded within undergrad and graduate level degree programs. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within undergrad and graduate level degree programs and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all undergrad and graduate level degree programs, accredited with professional qualifications, and an established part of continuing professional development.

3. Openness

Instruction: For this set of elements, consider the degree to which you engage in the following:

Data Table 3. CCMF- Openness items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
3.1 Openness in the course of research	No sharing. No details released.	Selected details released, e.g. in a proposal or project plan.	Selected intermediate results are shared within a limited group (besides mandated reporting to funders).	Intermediate results are shared through traditional means, e.g. conference papers.	Sharing is done publicly on the web. Full details are disclosed.
3.2 Openness of published literature	No sharing of papers or metadata outside publication channels.	Authors share metadata for their publications (e.g., abstracts, annotated citations)	Authors share theses or other selected sections from the literature.	Authors provide copies of their publications on request or other negotiated means.	Publications are made available on open access (e.g., repositories such as e-Pubs or public websites)
3.3 Openness of data	No sharing. No details released.	The data are described in the literature but not made available.	Data are available on request, after embargo or with other conditions.	Efforts are made to make data discoverable and re-usable as well as available.	Data is available in re-usable form and freely available to all. Community curation of the data may be possible.
3.4 Openness of research methodologies and workflows (e.g steps for preparing an interview or a focus group, how to run different statistical models on a software program)	No sharing. No details released	Released within limited scope.	Only partial stages of the workflow are openly shared.	The details of the workflow are shared but not the underlying scripts.	Sharing publicly on the web. Non-standard scripts, tools and software released.
3.5 Reuse of existing data and materials (including secondary sources, government statistics, photos in others' books)	Only own data or materials used.	Data exchanged within limited scope e.g. with collaborators or personal contacts	Use of data from repositories or other third parties.	Regularly combine data sets in specific established ways. Provenance tracked in ad hoc ways.	Multiple existing datasets often combined. Provenance tracked systematically.

4. Technical Infrastructure

Instruction: For this set of elements, describe the degree to which tools, infrastructure or support exists: If you know of specific tools, infrastructure or support you use, etc. please provide it on the comment section.

Data Table 4. Technical infrastructure items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
4.1 Computational tools and algorithms	None, home-grown or unknown	Tools exist but perform below requirements	Tools need to be customized for specific use-cases.	Tools have sufficient features to meet the needs of most users.	Tools have features expected to meet users' needs for the next few years
4.2 Tool support for data capture and collection (e.g., Screencasting tools, digital audio recorder, Web content scripters, Qualtrics, SurveyMonkey)	None, home-grown or unknown	Tools do not meet user requirements well or do not interoperate. Tools are custom and quality varies.	One or two good tools available. A few clear leaders	Most tools that support data capture do it well and meet user requirements	All tools support data capture well and interoperate. There is a good choice of tools for data processing
4.3 Tool support for data processing and analysis (e.g., Speech recognition/transcription tools, NVivo, ATLAS.ti, audio editors such as audacity)	None, home-grown or generic, not customized for your workflows	Tools do not meet user requirements well or do not interoperate. Tools are custom and quality varies.	One or two good tools available. A few clear leaders	Most tools that support data capture do it well and meet user requirements	All tools support data capture well and interoperate. There is a good choice of tools for data processing
4.4 Data storage	None, home-grown or unknown	Insufficient data storage available to meet user needs.	Although data storage is sufficient, tools do not interoperate e.g., no desktop tools to facilitate upload, versioning, etc.	Dedicated storage facilities are well integrated with other tools e.g., desktop tools to facilitate upload, versioning, etc. are in use	Storage is available and is expected to meet future needs
4.5 Support for data preparation for preservation (e.g., workflow to prepare data in repositories or data centers)	None, home-grown or unknown	Support is only available in specialized cases	Insufficient tools and facilities exist to meet needs.	Dedicated tools are available and are widely used	Common infrastructure is well funded and well used

4.6 Data/material discovery and access	None, home-grown or unknown	Discovery and access restricted to collaborators or personal contacts e.g. departmental or project intranet	Discovery services very discipline-specific; require specialized knowledge or rights e.g. PubMed	Discovery opened to all but siloed (not interoperable or easy to customize e.g. Dropbox)	Data discoverable and accessible to all, good integrated services
4.7 Integration and collaboration platforms or portal	None, home-grown or unknown	Platforms exist but perform below requirements.	Platforms need to be customized for specific use-cases.	Platforms have sufficient features to meet the needs of most users.	Platforms have features few people use, expected to meet users' needs for the next few years.
4.8 Data visualizations and representations (e.g., Using data to create visualizations)	None, home-grown or unknown	Tools exist but perform below requirements.	Tools need to be customized for specific use-cases.	Tools have sufficient features to meet the needs of most users.	Tools have features few people use, expected to meet users' needs for the next few years.
4.9 Platforms for citizen science (e.g., eBird, Old Weather)	None, home-grown or unknown	Tools built for individual use-cases.	Customized tools available, used by a small number of groups	Very flexible tools available and well used	Tools have been re-deployed to other disciplines.

5. Common Practices

Instruction: For this set of elements, consider the degree to which you adhere to the following:

Data Table 5. CCMF- Common practices items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
5.1 Data formats (e.g. The way that research data are stored and shared, such as MP3 for an audio file, TXT for text)	No standard formats available: ad hoc formats proliferate.	Standard formats are in development but not yet in use.	Some standard formats available but not widely adopted or community begins to converge on small number of formats.	Standard formats are widely adopted for some but not all types of data.	Standard formats are universally adopted for all types of data. Faithful conversions are possible between 'rival' standards.
5.2 Data collection methods (including sampling methods)	Methods are not usually shared.	Methods are shared but not widely reused.	Agreed methods are in development.	Although some methods are agreed there are gaps in the methods covered or room for improvement in the quality.	Methods are well known, well documented and well used.
5.3 Data processing workflows (i.e. systemized or automated workflow for processing samples, transcribing data, cleaning dataset, etc.)	Workflows are not usually shared.	Workflows are shared but not widely reused.	Agreed workflows are in development, or community begins to converge on a small number of workflows.	Agreed workflows are available with some gaps, or room for improvement in quality.	Several standardized workflows widely used.
5.4 Data description	No standard metadata schemes exist.	Standard metadata schemes are in development but not yet in use.	Some metadata schemes are published and recognized, but with little uptake or known flaws.	Recognized metadata schemes agreed, with some gaps.	Mature, agreed and widely used metadata schemes exist.
5.5 Standard vocabularies (e.g., American Folklore Society Ethnographic Thesaurus), semantics, ontologies	No standard schemes are available.	Some schemes are published but they are experimental with limited uptake.	Standards are being actively developed; agreement and standardization by the community is being pursued.	Some standard schemes are available, however gaps still exist.	Standard schemes are mature with good take-up by the community and widely applied.

5.6 Data identifiers such as Digital Object Identifiers etc, which are used to uniquely identify an object on the Web.	None in use.	Some used experimentally. Sporadic use.	Some trustworthy identifiers adopted.	Discipline-specific identifiers widely used.	International, well managed, sustainable schemes routinely used.
5.7 Stable, documented APIs (Application Programming Interface—examples for APIs: WorldCat Search API, Google Maps APIs, Twitter APIs, or government-related APIs which allows data harvesting from government data)	APIs not generally published or used.	Some tools offer APIs but with insufficient documentation.	A handful of well recognized APIs but these are the exception rather than the norm.	Most key disciplinary tools and services have useful, stable, and documented APIs.	Culture of developing APIs widespread.
5.8 Data packaging and transfer protocols (i.e., for record conversion, file compression, etc.)	Packaging and transfer performed ad hoc.	Standard protocols are in development but not yet in use.	Some standard protocols available but not widely adopted or community begins to converge on small number of protocols.	Some standard protocols available with some gaps, or room for improvement in quality	One or two standardized formats/protocols widely used

6. Economic & Business models

Instruction: For this set of elements, consider the scope and/or level of funding for the majority of your research:

Data Table 6. CCMF- Economic and business models items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
6.1 Duration of funding for research	Instruction: One-off funding focused on quick returns e.g. 1-2 years	Funding focused on short-term projects and quick returns e.g.2-3 years	Longer term investments on a 3-5 year timescale.	Single-phase thematic investments on a 5-7 year timescale.	Multi-phase thematic investments in 5-10 year blocks which build a community e.g. NSF DataONE Program
6.2 Geographic scale of funding for research	Projects funded internally.	Projects funded through grants from regional agencies.	Projects funded by national funders.	Projects funded by multiple national funders	Funding by international bodies and bi-lateral initiatives between national funders.
6.3 Scale of research that funding allows	Short investigative projects to encourage open innovation, usually conducted by a single scholar or team of 2	Small-scale projects (e.g., 3-5 scholars involved)	Mid-scale projects (e.g., 5-10 scholars involved)	Major investment (e.g., 10-20 scholars involved)	Large multi-national projects, more than 20 scholars collaborated e.g. EU's ERPANET (Electronic Resource Preservation and Access Network)
6.4 Sustainability of funding for infrastructure (i.e., building core network, IT services, and applications.)	One-off investments with no commitment to sustainment e.g. funding for start-up equipment: camera, digital audio recorder, tablets etc.	Multi-phase projects to develop infrastructure e.g. networks and services	Sustained multi-decade investments in data centers and services.	Infrastructure projects allowed slow transition to self-financing model.	Self financing infrastructure, networks and services
6.5 Geographic scale of funding for infrastructure (i.e.,	Projects funded internally (e.g., within a	Investments by a single funding body	Investments by a single funding body	Collaborative development at the national	Collaborative development between

building core network, IT services, and applications.)	department or an institution)	at regional level (e.g., city and state level)	at national level.	level by multiple funders	international funders
6.6 Scale of infrastructure (i.e., building core network, IT services, and applications.) projects that funding allows	Small-scale tool development (e.g. student built tools/ applications/instruments)	Medium scale investments in network services and systems e.g. Institutional Repositories	Co-ordinated investments at a regional level e.g. regional cloud services	Large central investments in network infrastructure or tools at a national level	Large multi-national investments which join multiple data centers
6.7 Public–private partnerships	None or unknown.	Informal collaboration with industry but no funding involved.	Corporate non-funded partners in proposals with academia e.g. through support letters, endorsements, MOUs etc.	Research is co-funded by industry and other sources.	Established formal co-investment partnerships running long-term multi-phase projects.
6.8 Productivity and return on investment	Long lead times between project start and submission of outputs (e.g. 6 years), and between acceptance and publication of papers (e.g. 2 years).	Long-mid range lead times between project start and submission of outputs (e.g. 4 years), and between acceptance and publication of papers (e.g. 18 months).	Mid-range lead times between project start and submission of outputs (e.g. 3 years), and between acceptance and publication of papers (e.g. 1 year).	Mid-short range lead times between project start and submission of outputs (e.g. 2 years), and between acceptance and publication of papers (e.g. 6 months).	Short lead times between project start and submission of outputs (e.g. 1 year), and between acceptance and publication of papers (e.g. 3 months).

7. Legal, Ethical & Commercial Issues

Instruction: For this set of elements, consider the extent to which these issues apply to, or are addressed in, your research:

Data Table 7. CCMF- Legal, ethical & commercial issues items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
7.1 Legal and regulatory frameworks e.g. IRB, related to sensitive data, patient records, human subjects, especially special classes of subjects (e.g., children or prisoner) etc	No coordinated response to legal, regulatory and policy issues. Confusion over obligations is widespread.	Basic frameworks exist but they are disjointed and frequently more hindrance than help.	Moderately sophisticated and helpful frameworks exist, but awareness of them is poor and the corresponding procedures are not well enforced.	Robust frameworks and procedures exist and are regulated at institutional level, but researchers do not fully trust them.	Trusted frameworks and procedures are in place. Discipline is well regulated by disciplinary bodies, professional societies.
7.2 Management of ethical responsibilities and norms e.g. Responsible Conduct of Research (RCR)	No standard procedures in place. Poor or uneven awareness of ethical issues and how to approach them.	Some procedures exist but they lack consistency, may hinder rather than help, and are rarely followed.	Consistent and useful procedures exist but they are not enforced.	Robust procedures are in place and are enforced locally, though they may be seen as a burden.	Trusted and accepted procedures are in place, and are enforced at the national or international level.
7.3 Management of commercial constraints e.g. as relates to intellectual property, copyright, patents, etc.	No standard procedures in place. Poor or uneven awareness of commercial issues and how to approach them.	Some procedures exist but they lack consistency.	Consistent and useful procedures exist but they are not enforced.	Robust procedures are in place and are enforced locally, though they may be seen as a burden.	Trusted and accepted procedures are in place, and are enforced at the national or international level.

8. Research Culture

Instruction: For this set of elements, consider the degree to which they apply to the environment in which you do research:

Data Table 8. CCMF- Research culture items

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
8.1 Entrepreneurship, innovation and risk	Highly risk-averse	Moderately risk averse	Calculated risks taken	Moderately innovative and experimental or exploratory with no certain outcome	Highly innovative and experimental
8.2 Reward models for researchers e.g. awards and other recognition besides tenure	None available	Narrow range of contributions recognized.	Wider range of contributions recognized, but informally.	Measures exist for more than one type of contribution and are well recognized.	All contributions are recognized and rewarded, through established procedures and measures.

APPENDIX C. LIST OF SAMPLED SOCIAL SCIENCE RELATED UNITS

University of Pittsburgh:

- Dietrich School of Arts and Sciences
    - Department of Communication
    - Department of Economics
    - Department of History
    - Department of History of Art and Architecture
    - Department of History and Philosophy of Science
    - Department of Political Science
    - Department of Psychology
    - Department of Sociology
- School of Education
    - Department of Administrative and Policy Studies
    - Department of Health and Physical Activity
    - Department of Instruction and Learning
    - Department of Psychology in Education
- School of Information Sciences
    - Department of Library and Information Science
- School of Law
- Graduate School of Public and International Affairs
    - Public Administration
    - Public & International Affairs
    - International Development
    - Public Policy & Management
- School of Social Work

Carnegie Mellon University

- Dietrich College of Humanities and Social Sciences
    - Department of History
    - Department of Psychology
    - Department of Social and Decision Sciences
- Heinz College
    - School of Public Policy & Management

APPENDIX D. PRELIMINARY INSTRUMENT ITEM SUMMARY

Data Table 9. Preliminary instrument summary

Dimensions	Attributes	# of items
Data Characteristics	DC1. User of data	8
	DC2. Data source	7
	DC3. Data types	
	DC4. Data volume	3
	DC5. Data sensitivity	
	DC6. Data's shareability	1
	DC7. Data ownership	1
Technical Infrastructure	TI1. Platform availability	3
	TI2. Platform usability*	0
	TI3. Facilities	6
	TI4. Technical standards*	0
Organizational and Research Context	OC1. Funding sufficiency	1
	OC2. Research data service (RDS) supports	3
	OC3. Internal human resources	7
	OC4. Legal and policy	1
	RC1. Discipline culture	6
	RC2. Discipline norms	2
	RC3. Research skills	9
	RC4. Research activities	11
Individual Characteristics and Motivations	IC1. Researchers' demographics	8
	IC2. Cost effectiveness	5
	IM1. Extrinsic motivation	3
	IM2. Scholarly Altruistism	2
Research Product Sharing Practices	DS1. Data sharing (channels and frequencies)	6
	DS2. Manuscript sharing (channels and frequencies)	5
Open-ended	comments	1

APPENDIX E. INSTRUMENT 2

Dear Madam or Sir,

As mentioned in the email, you are invited to participate in a study about social scientists' data-sharing experiences because you have deposited your data or been marked as a contributor at the {ICPSR or Qualitative Data Repository (QDR)}. In this study, we are especially curious about how social scientists prepare or share their data generated from qualitative or mixed-methods. The project is sponsored by Andrew W. Mellon Foundations and reports to the Council on Library and Information Resources (CLIR).

The outcome of this survey is expected to gain more insights on social scientists' actual practices on their qualitative data, and help us later to propose a more realistic data sharing guidelines that features discipline culture and academic norms to help researchers in social science handle their research materials.

The study project has been reviewed by the Institutional Review Board at University of Pittsburgh and meets all the necessary criteria for an exemption (IRB#: PRO15050056). Before agreeing to participate, please take two minutes to read:

Survey content and the estimated time

If you are willing to participate, we will ask about your background, experience, and thoughts about

research data sharing. If you have experience on qualitative data sharing, we will have a few further questions. According to our pre-test results, the estimated completion time of this survey is approximately 8-10 minutes.

Voluntary participation and right to withdraw

Your participation in this study is voluntary, and you may stop completing the survey . If there is an early withdrawal from the study, the data will not be included in the analysis and be destroyed immediately.

Research data confidentiality and data sharing

Your responses and any personal information, such as research background, will remain confidential during the course of research process. Your contact email will not be linked to your response in any case. Any personal information that could identify you as an individual will be removed or changed before data are shared with other researchers or research findings are made public.

Contact information and the research team

This study is conducted by the project manager, Wei Jeng, PhD candidate, and advised by Dr. Daqing He, associate professor, both in the School of Information Sciences at the University of Pittsburgh, who can be reached at wej9@pitt.edu and dah44@pitt.edu if you have any questions.

If you agree to participate AND identify yourself as a social scientist, please check "Yes" and click "Next."

- ❑ Yes (1)

Section 1. Research Background

Please answer the following questions about the characteristics of your research and your data.

Which of these best describe your primary subject discipline?

- ❍ Anthropology (1)
- ❍ Archeology (2)
- ❍ Area/Ethnic/Cultural/Gender Studies (3)
- ❍ Business, Management & Administration (4)
- ❍ Communication Research (5)
- ❍ Criminology/Criminal Justice (6)
- ❍ Economics (7)
- ❍ Education (8)
- ❍ Family/Consumer Science/Human Science (9)
- ❍ Geography (10)
- ❍ History (11)
- ❍ Humanities, Other (12)
- ❍ International Relations/Affairs (13)
- ❍ Law (14)
- ❍ Library Science/Information Science/Archival studies (15)
- ❍ Linguistics (16)
- ❍ Philosophy (17)
- ❍ Political Science & Government (18)
- ❍ Public Administration (19)
- ❍ Public Policy Analysis (20)
- ❍ Psychology (21)
- ❍ Social Sciences, Other (22)
- ❍ Social Work (23)
- ❍ Sociology (24)
- ❍ Statistics (25)
- ❍ Other (please specify in the below text box) (99) _____

Please briefly specify your area(s) of research interest by providing some keywords. (open-ended)

Are your research interests currently joined with other discipline(s)?

- ❍ Yes, please list the name of your secondary field. (1) _____
- ❍ No (2)

Which of these describe the type(s) of data you usually interact with in your research career?

The definition of research data in social sciences is "materials generated or collected during the course of conducting research."

- ❑ Observational data captured in real time (e.g., fieldnotes, social experiments) (1)
- ❑ Data directly obtained from the study groups/informants (e.g., survey responses, diaries, interviews, oral histories) (2)
- ❑ Experimental data (e.g., log data) (3)
- ❑ Simulation data generated from test models, where models are more important than output data (e.g., economic models) (4)
- ❑ Records, literature, archives, or other documentation (e.g., court records, prison records, letters, published articles, historical archives) (5)
- ❑ Secondary data (e.g., government statistics, data from IGOs or NGOs, other's data) (6)
- ❑ Physical materials (e.g., artifacts, samples) (7)
- ❑ Other (please specify) (99) _____

Please recall one of your most recent research projects and estimate the proportion of your qualitative (QUAL) data, compared with your quantitative (QUANT) data in it.

The qualitative data are data generated from qualitative approaches or involved qualitative judgments, such as interviews, open-ended surveys, focus groups, oral histories, observations, or content analysis.

- ❍ Purely QUANT data (1)
- ❍ Mix, with more QUANT data (2)
- ❍ About an equal mix of both (3)
- ❍ Mix, with more QUAL data (4)
- ❍ Purely QUAL data (5)

Section 2. Data-sharing practices

Please answer the following questions about your experiences and attitudes regarding research data sharing.

Data sharing means providing the raw data of your research project to other researchers outside of your research team(s) by making it accessible through data repositories, public web space, social media, publications' supplementary materials, or by sending the data via personal communication methods upon request.

Display Logic:
If in question "please estimate the proportion of your qualitative data (QUAL), compared with your quantitative data…", Purely QUANT data is not selected

Based on your overall experience, which data or materials at below would you be willing to share with other researchers?

	Very Unlikely (1)	Somewhat Unlikely (2)	Neutral (3)	Somewhat Likely (4)	Very Likely (5)	I don't usually handle this kind of data (99)
Procedures of data collection e.g., a focus group protocol	❍	❍	❍	❍	❍	❑
Researchers' notes	❍	❍	❍	❍	❍	❑
Survey/ interview instruments with actual questions	❍	❍	❍	❍	❍	❑
Analysis data/scripts such as qualitative data analysis software files e.g., files on NVivo, ATLAS.ti	❍	❍	❍	❍	❍	❑
Individual survey responses	❍	❍	❍	❍	❍	❑
Interview transcripts	❍	❍	❍	❍	❍	❑
Multimedia files related to study	❍	❍	❍	❍	❍	❑

In the past five years, how frequently have you shared or deposited the data for your research project(s) through these channels?

1. Never or Rarely (about 0-10% of the time)

2. Occasionally (about 25% of the time)

3. Sometimes (about 50% of the time)

4. Often (about 75% of the time)

5. Frequently or Always (about 90-100% of the time)

	Never or Rarely (1)	Occasionally (2)	Sometimes (3)	Often (4)	Frequently or Always (5)
Institutional repositories	❍	❍	❍	❍	❍
Public Web spaces (e.g., your website)	❍	❍	❍	❍	❍
Academic social media platforms (e.g., ResearchGate, figShare)	❍	❍	❍	❍	❍
Discipline data repositories (e.g., ICPSR, QDR)	❍	❍	❍	❍	❍
Via emails (e.g., after receiving a direct request from other researchers)	❍	❍	❍	❍	❍
Publications as supplemental materials	❍	❍	❍	❍	❍

How much do you agree with the following statements in terms of the factors that might influence your decision to share data?

I will be more willing to share data if...

	Strongly Disagree (1)	Somewhat Disagree (2)	Neither agree or disagree (3)	Somewhat Agree (4)	Strongly Agree (5)
I have complete rights to make the data public.	❍	❍	❍	❍	❍
the ownership of my research data completely belongs to me.	❍	❍	❍	❍	❍
my data is interpreted in an appropriate way.	❍	❍	❍	❍	❍
my data is re-used in an appropriate way.	❍	❍	❍	❍	❍
I have confidence in the overall data quality in my research (e.g., few errors).	❍	❍	❍	❍	❍
I have confidence in the strength of evidence that I use in my research.	❍	❍	❍	❍	❍

Section 3. Discipline Community and Perceived Technological Supports

Please answer the following questions about your discipline community and work environment regarding research data sharing.

To what degree do you agree with the following statements describing your discipline community in terms of data sharing?

In my discipline community,

	Strongly Disagree (1)	Somewhat Disagree (2)	Neither Agree or Disagree (3)	Somewhat Agree (4)	Strongly Agree (5)
it's common to see people sharing their data.	❍	❍	❍	❍	❍
people care a great deal about data sharing.	❍	❍	❍	❍	❍
there is a generic standard for data sharing.	❍	❍	❍	❍	❍

Based on your past impressions, please rate the technology related resources that exist in your work environment.

In my work environment, technology related to...

	Very Insufficient (1)	Somewhat Insufficient (2)	Moderate (3)	Somewhat Sufficient (4)	Very Sufficient (5)
collecting data	❍	❍	❍	❍	❍
analyzing data	❍	❍	❍	❍	❍
helping researchers to discover others' data	❍	❍	❍	❍	❍
helping researchers prepare data for sharing	❍	❍	❍	❍	❍

The following statements relate to your thoughts about sharing data with others. Please tell us how much you agree with the following statements. Data sharing can...

	Strongly Disagree (1)	Somewhat Disagree (2)	Neither Agree or Disagree (3)	Somewhat Agree (4)	Strongly Agree (5)
help my publications earn more citations.	❍	❍	❍	❍	❍

help advance my career.	❍	❍	❍	❍	❍
give me an opportunity to collaborate with other researchers.	❍	❍	❍	❍	❍
help others to fulfill their research need.	❍	❍	❍	❍	❍
provide a sample for others to learn about practicing social research methods.	❍	❍	❍	❍	❍
inspire other researchers or students.	❍	❍	❍	❍	❍

Given the following conditions, how likely are you to share your data with others? I am willing to share my data if ...

	Strongly Disagree (1)	Somewhat Disagree (2)	Neither Agree or Disagree (3)	Somewhat Agree (4)	Strongly Agree (5)
I have sufficient time.	❍	❍	❍	❍	❍
a small amount of effort is required.	❍	❍	❍	❍	❍
I have sufficient funds for the data deposit fee.	❍	❍	❍	❍	❍
it's easy to find an appropriate place to deposit my data.	❍	❍	❍	❍	❍
I have a better sense of good practices in data sharing.	❍	❍	❍	❍	❍

Section 4. Demographics

Which one of the following best describes your primary work sector?

- ❍ Academic (1)
- ❍ Government (2)
- ❍ Non-profit (3)
- ❍ Commercial / Industrial (4)
- ❍ Other (please briefly specify:) (5) _____

Which one of the following best describes your current position?

- ❍ Professor (1)
- ❍ Associate professor (2)
- ❍ Assistant professor (3)
- ❍ Researcher associate / scientist (4)
- ❍ Post-doctoral researcher (5)
- ❍ Graduate student (6)
- ❍ Administrator (7)
- ❍ Professor emeritus (8)
- ❍ Other (please briefly specify:) (99) _____

Which one of the following best identifies your gender?

- ❍ Female (1)
- ❍ Male (2)
- ❍ Prefer not to answer (99)

Your age group:

- ❍ 18-34 (1)
- ❍ 35-44 (2)
- ❍ 45-54 (3)
- ❍ 55-64 (4)
- ❍ 65+ (5)
- ❍ Prefer not to answer (99)

Any comments before your submission? Please feel free to use this space and write down your thoughts and comments regarding research data sharing in general or regarding this project. (open-ended question)

APPENDIX F. SUPPLEMENTAL DATA TABLES IN CASE STUDY 1 AND CASE STUDY 2

Data Table 10. Demographic of participants

		Case 1 (n=66)		Case 2 (n=70)	
		N	%	N	%
Position	Full rank professor	0	0	29	41.4%
	Associate professor	0	0	13	18.6%
	Assistant professor	0	0	1	1.4%
	Research associate/fellow	0	0	11	15.7%
	Post-doctoral researcher	1	1.5%	1	1.4%
	Graduate student	62	94%	3	4.3%
	Administrator	0	0	6	8.6%
	Professor emeritus	0	0	2	2.9%
	Other	2	3%	4	5.7%
Gender	Female	38	57.6%	26	37.1%
	Male	27	40.9%	44	62.9%
Age group	18-24	6	9.1%		
	25-34	50	75.8%	1*	1.4%
	35-44	5	7.6%	13	18.6%
	45-54	1	1.5%	30	42.9%
	55-64	3*	4.5%	14	20.0%
	65+			12	17.1%
Discipline group*	Economics & Business	10	15.2	3	4.3
	Education	11	16.7	1	1.4
	Geography			1	1.4
	History	2	3	--	--
	Info and Communication	9	13.6	2	2.9
	Law, Criminology & Criminal Justice	--	--	12	17.1
	Political, Government & Policy	15	22.7	16	22.9
	Psychology & Decision Sciences	12	18.2	9	12.9
	Public health & Family Studies	1	1.5	11	15.7
	Sociology & Social Work	6	9.1	11	15.7
	Social Sciences, General	--	--	4	5.7
Work sector	Academic	66*	100%	60	84.5%
	Goverment			2	2.8%
	Non-profit			7	9.9%
	Commercial or industrial			1	1.4%
	Other			1	1.4%

Data Table 11. Raw data of discipline

Discipline	Case 1		Case 2	
	N	%	N	%
Business, Management & Administration	5	7.6%	1	1.4%
Communication Research	2	3.0%	--	--
Criminology & Criminal justice	--	--	11	15.7%
Economics	5	7.6%	2	2.9%
Education	11	16.7%	1	1.4%
History	2	3.0%		
Family/ Consumer Science/ Human Science	--	--	1	1.4%
Geography	--	--	1	1.4%
International Relations/Affairs	2	3.0%	1	1.4%
Law	0	0%	2	2.9%
Library Science/Information Science/Archival studies	6	9.1%	--	--
Philosophy	--	--	1	1.4%
Political Science & Government	7	10.6%	13	18.6
Public Policy Analysis	4	6.1%	1	1.4%
Social Work	1	1.5%	1	1.4%
Sociology	4	6.1%	9	12.9
Statistics	3	4.5%	1	1.4%
Psychology	10	15.2%	9	12.9
Decision making	2	3.0%	--	--
Social Psychology & Social Network	2	3.0%	--	--
Other	--	--	15	21.4%

Data Table 12. Data sources in Case Study 1 and 2

Data source	Case 1 (n=66)		Case 2 (n=70)	
	N	%	N	%
Data from observational studies	32	48.5%	32	45.7%
Directly from informants	48	72.7%	65	92.9%
Data from experimental studies	19	28.8%	20	28.6%
Data from simulation	9	13.6%	8	11.4%
Records	21	31.8%	31	44.3%
Secondary	34	51.5%	54	77.1%
Materials	--	--	4	5.7%

Data Table 13. Cross-tabulation of discipline and preferred research methods

Discipline	Pure QUAL	Mix but QUAL more	Equal Mix	Mix but QUANT more	QUAN	TOTAL
Business, Management & Administration	0	0	0	0	1	1
Criminology & Criminal Justice	1	2	1	4	3	11
Economics	0	0	0	1	1	2
Education	0	1	0	0	0	1
Family/Consumer Science/Human Science	0	0	0	1	0	1
Geography	0	0	1	0	0	1
Law	0	0	1	0	0	1
Library Science/Information Science/Archival studies	0	0	1	1	0	2
Philosophy	0	0	0	1	0	1
Political Science & Government	2	2	1	5	3	13
Psychology	0	0	2	5	2	9
Public Policy Analysis	0	0	0	0	1	1
Social Sciences, Other	0	0	0	4	1	5
Social Work	0	0	0	1	0	1
Sociology	0	0	2	1	6	9
Statistics	0	0	0	0	1	1
Other	0	0	1	6	3	10
Total	3	5	10	30	22	70

Data Table 14. Cross-tabulation of discipline and proportion

	Purely QUANT	QUANT more	Equal	QUAL more	Purely QUAL	Total
Economics & Business	2	1	0	0	0	3
Education	0	0	0	1	0	1
Geography	0	0	1	0	0	1
Info and Communication	0	1	1	0	0	2
Law, Criminology & Criminal Justice	3	4	2	2	1	12
Political, Government & Policy	4	7	1	2	2	16
Psychology & decision making	2	5	2	0	0	9
Public health & Family	2	8	1	0	0	11
Social Sciences, General	2	2	0	0	0	4
Sociology & Social Work	7	2	2	0	0	11
Total	22	30	10	5	3	70

APPENDIX G. FOCUS GROUP INFORMED CONSENT IN CASE STUDY 3

Below text was both presented and read to all focus group participants in CS3:

The purpose of this study is to understand data curators' practical experiences on curating and developing collection on social science data.

To achieve this goal, we will be conducting two focus groups in the world's largest social science data repository- The Interuniversity Consortium for Political and Social Research (ICPSR). Participants include professionals and managers in data curation services and collection development. The participants' views and practical experiences on curating data, developing collections, and managing a data repository will be used to help us understand practitioners and curators' views on social science data reuse and sharing, especially for mixed methods and qualitative data. All participants must be 18 years of age or older.

If you are willing to participate, we will ask about your education background, your practices at ICPSR, and experiences and insights related to your responsibilities at ICPSR. The estimated completion time is approximately 60 minutes.

Please note that *audio recording without facial features will be taking place* for only transcribing purposes. Recordings will not be made to public. You can request to turn off the recording any time. The results will be stored in password-protected computers. There are no foreseeable risks associated with this project, nor are there any direct benefits to you.

Our study project has been reviewed by the Institutional Review Board at University of Pittsburgh and meets all the necessary criteria for an exemption (IRB#: PRO15050056). According to the Basic Exempt Criteria 45 CFR 46.101(b)(2), we are allowed to obtain every participant's oral agreement, but no formal written consent is obtained.

The information in this study will be used only for research purposes and in ways that will NOT reveal who you are. Federal or state laws may require us to show information to university or government officials, or sponsors, who are responsible for monitoring the safety of this study. However, an assigned participant number will be used to designate your record with your responses and not information that personally identifies you. Any personal information that could identify you will be removed or adjusted before result are revealed in any way, including publishing, sharing with other researchers, or making datasets to public.

Your participation is voluntary, and you may stop completing the interview at any time.

This study is being conducted by Wei Jeng, with Yu Chi, and Daqing He. You could also contact the PI, Wei Jeng, at wej9@pitt.edu for more questions about this study.

APPENDIX H. FOCUS GROUP PROTOCOLS

Group A (data curation professionals): 60 minutes

Data Table 15. Protocol for Group A

Time	Activity	Mediator actions	Question prompts
00:00-00:03	Review information and consent	Distribute introduction script Obain consents on: ▪ proceed the focus group ▪ use recorders, and ▪ data will be shared	Thank you for your participation. I believe your input will be valuable to this research and in helping grow all of our professional practice. Approximate length of interview: 60 minutes, two group activities and three major questions
00:03-00:15	Warming up	Mediator actions ▪ Set timer ▪ Set recorder Taking note: ▪ Education background ▪ Career history ▪ Year of experience ▪ Primary activities	Please take us back through a little history in your career that brought you to this current position. Also, we would like to know more about your current work at ICPSR. Prompts: ▪ How long have you been involved in your current job? (What year were you involved) ▪ What primary tasks does your job involve?
00:15-00:35	Concept construction	Distribute post-its (different colors) Process: individual write post-its stick to write board sort cluster 📷 Take a picture Distribute easel pad 📷 Take a picture Distribute post-its (yellow post-its) 📷 Take a picture	**Question 1**: What are your activities as a curation professional to support data curation? Prompt: before/ after data submitting Process: individual write post-its→ stick to write board → sort→ draw cluster→ ask participants if there is anything left. **Question 2**: Now we have n clusters, could you explaining the relationships among the activities

			Question 3: What are the tools that you use for your actions in curation? Prompts: ▪ Computer equipments ▪ Software ▪ Online services ▪ Internal toolkits? Question 4: Can you think of any desired tools or technology (tools may not exist) which can facilitating your actions at ICPSR? (talking only, do not distribute sticky notes)
00:40-00:55	Questions about qualitative data curation	--	**Question 5A**: Have you ever curated qualitative data? If yes, jump to 5B If no, have you heard about your colleagues or others in ICPSR curating qualitative data? Do you have any observation? **Question 5B:** Please tell us about the difference when curating qualitative, mixed method, and quantitative data, if any. Is there any special case or example that you would like to share? **Question 6:** Based on your observations and experience as curation professionals in ICPSR, what are the critical factors that may influence a PI's willingness to share his/her data? Prompts: ▪ Has a PI ever told you about or you have heard---the factors could influence PI's willingness? ▪ Are they from: ▪ Individual incentives ▪ Research culture ▪ Institution
00:55-00:60	Debriefing	--	Suggestions about research instrument? Was anything unclear?

Group B (collection development professionals): 60 minutes

Data Table 16. Protocol for Group B

Time	Activity	Mediator actions	Question prompts
00:00-00:03	Review information and consent	Distribute introduction script Obain consents on: ▪ proceed the focus gorup ▪ use recorders, and ▪ data will be shared	Thank you for your participation. I believe your input will be valuable to this research and in helping grow all of our professional practice. Approximate length of interview: 60 minutes, two group activities and three major questions
00:03-00:15	Warming up	Mediator actions ▪ Set timer ▪ Set recorder Taking note: ▪ Education background ▪ Career history ▪ Year of experience Primary activities	Please take us back through a little history in your career that brought you to this current position. Also, we would like to know more about your current work at ICPSR. Prompts: How long have you been involved in your current job? (What year were you involved) What primary tasks does your job involve?
00:15-00:30	Concept construction	Distribute post-its (different colors) Process: individual write post-its stick to write board sort cluster 📷Take a picture Distribute easel pad 📷Take a picture Distribute post-its (yellow post-its) 📷Take a picture	**Question 1:** What are your responsibilities in supporting collection development and delivery in ICPSR? Prompt: before/ after data submitting Process: individual write post-its→ stick to write board → sort→ draw cluster→ ask participants to clarify if there is any sticky note unclassified. **Question 2**: Are there any tools that you use? Prompts: Computer equipments Software Online services Internal toolkits? (yellow post-its) Question 3: Can you think of any desired tools (tools may not exist) or technology which can facilitating your actions at ICPSR? (talking only)

00:30-00:55	Questions about collection development and vision		Now we have a couple questions related to collection development, collection delivery, management, and marketing topics in ICPSR. **Question 4:** How do you determine the scope of ICPSR's collection? We read about ICPSR's collection development policy, we read about the high-priority areas including sexual orientation, social media, immigration, and so on. How does ICPSR decide which areas should be given priority? Prompts: ▪ Are these decisions from ICPSR's interval decision? ▪ members' opinions or feedback? ▪ Recent research hot topics (recent publications)? ▪ or community or specific researchers' demands? ▪ How does ICPSR decide to add a new interest? **Question 5:** This question is related to appraisal standards in ICPSR. Please tell us about how ICPSR applies the selection and appraisal criteria for data from mixed-method study or qualitative study. Are they different from quantitative one? Is there any special case or example that you would like to share? Prompts: When will data be referred to the QDR? **Question 6:** This questions is about OpenICPSR. Given the differences between OpenICPSR and ICPSR, please share your experience with us about how ICPSR handles or manages these two different collections. Is OpenICPSR within the scope of ICPSR? Prompts: ▪ Do ICPSR members mention anything about ICPSR? (Their experience with OpenICPSR?) ▪ What is your observation? ▪ Is there any plan for further promoting Open-ICPSR to ICPSR members? **Question 7:** Currently ICPSR supports search interface and track utilization for data sharers and reusers. Does ICPSR provide other services or support to further connect the data depositors and reusers?
00:55-00:60	Debriefing	--	Suggestions about research instrument? Was anything unclear?